Machine learning for the prediction of infection and evaluation of maturation in premature infants combining cardiac and respiratory variability

Cristhyne Stephania Leon Borrego

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, Image, Vision*

Par
Cristhyne Stephania LEON BORREGO

Apprentissage automatique pour la prédiction de l'infection et de la maturation chez le grand prématuré en associant les variabilités cardiaques et respiratoires

Thèse présentée et soutenue à Rennes, le 12 juillet 2021
Unité de recherche : LTSI, UMR Inserm 1099 Laboratoire Traitement du Signal et de l'Image

Rapporteurs avant soutenance :

Julien DE JONCKHEERE	Ingénieur de Recherche, Inserm CIC-IT 1403 CHU Lille
Pablo LAGUNA	Professeur, BSICoS-I3A, Universidad de Zaragoza

Composition du Jury :

Président :	Catherine MARQUE	Professeur, BMBI-CNRS, Université de Technologie de Compiègne
Examinateurs :	Olivier BAUD	Professeur, Université de Genève, Suisse
	Patrick PLADYS	Professeur, LTSI-INSERM, Université de Rennes
Dir. de thèse :	Guy CARRAULT	Professeur, LTSI-INSERM, Université de Rennes

Acknowledgement

Like my meetings with my thesis supervisor, these acknowledgments will be in a flowing mix of languages, and although they might seem messy to some, they will make perfect sense and be packed with content.

Je veux commencer par remercier Guy. Je sais que tu aimes quand je parle français, alors je te remercierai en français. Je n'aurais pas pu rêver avec un meilleur directeur de thèse. Merci de m'avoir fait confiance, de m'avoir permis d'être créatif et de tester mes folles idées, mais aussi merci d'avoir eu la patience de battre avec moi quand j'étais têtue, ou testatruda, comme tu as appris à le dire.

Thanks to the Digi-NewB team, for welcoming into this project, for all the great work you had already done before I arrived, which made my research infinitely easier, for all the interesting discussions and exchanges, and for the heart and soul you continue to put into this project. Patrick, thank you for being the physician on call for all of my research's medical emergencies. Raphael and Edouard, thank you for being my welcome committee to the project and to the office. Sandie, thank you for your amazing work which so often motivated me to do mine better.

Gracias a Juan Pablo y a Gustavo, por ser mi familia en Francia. Juan, has sido ese hermano al que siempre puede ir con mis quejas y mis celebraciones; el hermano con el que casi siempre estoy de acuerdo, que siempre entiende, y que nunca juzga. Gus, tu has sido ese hermano con el que peleo constantemente, nunca estamos de acuerdo en nada, pero que al final del día sé que puedo contar contigo, hasta para que me hagas empanadas cuando tengo el corazón partido.

Thanks to all the amazing PhD students I have had the pleasure to meet at the LTSI, for all the wonderful lunches shared together (in those times before COVID that now seem so far away!) and all the fun outings. You made the work place fun, and the after work time lively. Special shout out to the original RU team: Kimi, Remo, and Gustavo.

Thanks, of course, to my International Refugees: Lorenzo, Genc, Younes, Paulo, Amir, and the honorary members Madalina and Bogdan. Thanks for your friendship, for being my welcome committee to Rennes, and for your hospitality. For all the international Mondays, the Saturday nights, and for offering me a place to go and feel like in family on Christmas Days.

Genc, faleminderit qe je zoti e Inkscape-it. Shumë faleminderit që mbush jetën time me ngjyra, me gjuhën tënde të mrekullueshme dhe fjalët shqip (dhe kulturën, dhe historinë, dhe këngët, dhe filmat, dhe librat, dhe plazhet, dhe ushqimin, dhe lista vazhdon), dhe për të shkuar në aventura me mua, të cilat shpesh i tejkaluan të gjitha të papriturat e Covid-it

Finalmente, gracias a mi familia por todo los esfuerzos y sacrificios que han hecho para que yo este aquí hoy, por todo el apoyo que me han brindado, y por ser constantemente mi fuente de inspiración. Especialmente gracias a Carine, por ser una de las personas mas guerreras y valientes que conozco, e inspirarme cada día a dar lo mejor de mi. Los amo.

Résumé en français

Les naissances prématurées, définies comme survenant avant 37 semaines d'âge gestationnel, représentent environ 10,6% des naissances dans le monde [1]. Leur prévalence en Europe varie de 6 à 12%, selon les pays. Les enfants nés prématurément présentent des taux de morbidité et de mortalité élevés, une durée d'hospitalisation plus longue et une incidence de réadmission plus élevée à l'hôpital après la première sortie [2, 3].

Il existe une grande variété de conditions qui affectent les nouveau-nés prématurés pendant la période postnatale. Beaucoup ont des conséquences à court et à long terme sur leur santé et leur développement. L'approche privilégiée pour améliorer l'état de santé de cette population particulièrement vulnérable consiste à diagnostiquer le plus précocement possible ces pathologies. Ceci explique que ces dernières décennies, un effort considérable a été mené pour intégrer des approches relevant de l'intelligence artificielle et de l'apprentissage automatique dans les systèmes de surveillance des nouveau-nés.

C'est dans cette optique qu'est né le projet Digi-NewB soutenu par la commission Union Européenne. Ce projet visait à proposer un nouveau système de surveillance pour les soins des prématurés. Son originalité reposait sur le fait qu'il avait pour ambition d'associer les mesures traditionnelles utilisées en unités de soins intensifs (signaux physiologiques, signes cliniques) à des nouvelles mesures jamais explorées ce jour en unités de soins intensifs telles que la video et les pleurs des bébés. Digi-NewB avait deux objectifs principaux : la proposition d'un système d'aide à la décision (DSS) pour la détection précoce de l'infection nosocomiale tardive, et on la dénotera dans la suite avec la notation anglaise Late Onset Sepsis (LOS), chez les prématurés et la proposition d'un système de surveillance pour la quantification de la maturation cardio-respiratoire et neuro-comportementale des prématurés pendant leur hospitalisation. Ce projet a été réalisé grâce à la collaboration de plusieurs partenaires publics et privés situés en Finlande, en France, en Irlande et au Portugal. Il a permis de recueillir des données sur plus de 600 prématurés dans six hôpitaux de la région ouest de la France.

Le travail décrit dans ce mémoire s'inscrit pleinement dans les objectifs de Digi-NewB et a été réalisé entièrement sur des données issues du projet. Cependant, il convient de préciser que cette thèse s'est concentrée exclusivement sur les applications fondées sur les signaux physiologiques, plus précisément sur les données de variabilité de la fréquence cardiaque, de variabilité de la fréquence respiratoire et les événements bradycardiques. Plus précisément, il s'agissait :

— D'introduire des nouveaux indices pour la caractérisation de la variabilité de la fréquence cardiaque et de mesurer leur impact sur les performances des modèles de détection du LOS.

— D'identifier les fenêtres temporelles optimales d'apprentissage des algorithmes d'apprentissage supervisé.

— De mettre en œuvre un modèle de détection précoce du LOS basé sur les données de variabilité de la fréquence cardiaque. Ce système a été pensé en tenant du fait qu'il devait être déployé en tant que système d'aide à la décision non invasif fonctionnant en temps réel dans une unité de soins intensifs néonatale (USIN).
— De proposer une métrique permettant d'évaluer objectivement le développement de la maturation des prématurés, sur la base de signaux physiologiques, pendant leur séjour à l'hôpital.
— D'exploiter ce modèle afin d'estimer et de détecter les perturbations de la maturation chez des prématurés présentant des troubles de développement.

Dans les sections suivantes, nous présentons les concepts de base exploités tout au long de la thèse et l'état de l'art sur lequel nos contributions ont été inspirées. Puis, nous détaillons les contributions apportées par cette thèse.

Les soins intelligents en néonatologie et la proposition Digi-NewB

L'intelligence artificielle est une branche de l'informatique dans laquelle s'inscrit l'apprentissage automatique ou *machine learning* en anglais. L'apprentissage automatique fait référence à la capacité des algorithmes à s'améliorer automatiquement en apprenant de l'expérience. Il existe de nombreux d'algorithmes capables d'apprendre et qui ont été développés ces dernières décennies. Parmi les plus courants et les plus utilisés, et qui ont été comparés dans ce travail, on peut citer : les K plus proches voisins (KNN), la régression linéaire, la régression logistique, les forêts aléatoires, les machines à vecteurs de support (SVM), les algorithmes génétiques et les réseaux neuronaux artificiels.

Bien que l'intelligence artificielle et l'apprentissage automatique ne soient pas des concepts nouveaux, ils ont connu un essor considérable ces dernières années avec l'arrivée des nouvelles technologies et d'Internet. En effet, ceci a rendu disponibles de grandes quantités de données permettant une généralisation à partir d'exemples différents. Dans le même temps, la néonatologie a profité de cet engouement puisque les unités de soins intensifs sont aussi capable de générer de grands volumes de données. On estime que plus 40 téraoctets de données par lit et par an sont produits [4].

Basés sur des algorithmes de *machine learning*, les premiers travaux ont visé donc à proposer des modèles pour la détection précoce du LOS en se basant sur une combinaison de signes cliniques, de résultats de tests de laboratoire et de signaux physiologiques. De très bonnes performances ont été atteintes. Toutefois, ces approches présentent l'inconvénient de reposer sur des données qui doivent être manuellement annotées par le personnel médical. Outre l'erreur humaine incontournable, ceci induit également un retard dans la prise de décision dans la mesure où l'on doit disposer des résultats des tests de laboratoire, qui plus est, sont invasifs. Ces constats

limitent de fait l'utilisation de ces modèles invasifs en temps réel. Une étude récente a proposé une solution pour contourner ces obstacles en exploitant uniquement les données de fréquence cardiaque [5]. Cette étude a montré que l'on pouvait réduire la mortalité des prématurés dans les USIN [6] en dépit de performances inférieures aux approches plus traditionnelles. Sur le plan décisionnel et du traitement du signal, ce travail s'appuie sur la régression logistique, reconnue comme un des algorithmes d'apprentissage automatique les plus simples, et exploite seulement trois caractéristiques de la variabilité de la fréquence cardiaque. Ces quelques limites soulignent donc tout l'intérêt de proposer une nouvelle approche non invasive pour la détection précoce du LOS en introduisant d'une part une comparaison avec des algorithmes plus récents du machine learning et d'autre part un plus grand nombre de caractéristiques liées à la variabilité cardiaque.

Comparée à la détection du LOS, la quantification de la maturation des prématurés par des applications d'apprentissage automatique a fait l'objet de moins de travaux. La plupart des études sur ce sujet se sont concentrées exclusivement sur l'évaluation directe de la maturation cérébrale à partir de données d'imagerie par résonance magnétique ou d'électroencéphalogramme. Certaines études ont suggéré l'estimation d'un âge de maturation cérébrale [7] ou fonctionnelle [8] à partir de ces données. Cependant, ces approches peuvent difficilement être applicables en USIN et on s'est interrogé si les données de fréquence cardiaque (HRV) ou de fréquence respiratoire (RRV), mesurées de manière continue en USIN, ne pourraient pas être exploites pour le suivi de la maturation des prématurés.

Le projet Digi-NewB visait donc à répondre à ces questions laissées sans réponse dans la littérature récente. En ce qui concerne le LOS, son objectif était de proposer un système d'aide à la décision (DSS) basé sur l'intégration de plusieurs signaux physiologiques, en améliorant à la fois les informations placées à l'entrée des algorithmes de machine learning et les algorithmes eux-mêmes. Digi-NewB visait également à fournir un système de suivi de la maturation des nouveaux nés sur le plan cardio-respiratoire et neurodéveloppemental en explorant d'autres sources de données, à ce jour jamais exploitées en USIN, qui peuvent être acquises de manière non invasives comme la vidéo et/ou les pleurs. L'objectif final était de proposer un système non invasif, sans nécessiter d'équipement supplémentaire en contact direct avec le prématuré (afin d'éviter toute perturbation supplémentaire de son environnement). Cependant, ces nouveaux systèmes de surveillance doivent également restés fiables, conviviaux et capables de fonctionner aussi près du temps réel que possible. C'est dans cet esprit que s'est déroulé ce travail et plusieurs contributions majeures ont été réalisées et sont résumées ci-après.

Intérêt des graphes de visibilité pour le diagnostic précoce de l'infection tardive chez les prématurés

Les indices dérivés des graphes de visibilité ont été proposés pour caractériser la variabilité cardiaque. Ils sont estimés à partir d'une transformation en réseau des séries temporelles de la variabilité de la fréquence cardiaque (HRV) [9]. Nous avons évalué l'intérêt d'introduire ces indices conjointement avec les caractéristiques du domaine temporel, des domaines fréquentiel et non linéaire qui sont généralement utilisées pour caractériser la HRV. Nous avons également testé i) l'impact sur les performances des modèles d'apprentissage automatique de l'utilisation de différentes périodes de calibration pour s'adapter à la HRV de base de chaque enfant, ii) différentes tailles de fenêtre pour l'apprentissage. Ce dernier problème est crucial puisque la date exacte d'infection du prématuré ne peut être connue. Pour ce faire, une population de 49 prématurés (24 infectés et 25 contrôles) ont été utilisés. Quatre algorithmes d'apprentissage automatique : le KNN, la régression logistique, les forêts aléatoires et les SVM ont été comparés pour cette étude.

Nous avons constaté que, sur les quatre algorithmes d'apprentissage automatique, trois ont obtenu les meilleurs résultats lorsque la période de calibration du comportement HRV de base de chaque prématuré avait une durée de 48 heures. Seul, l'algorithme SVM a obtenu de meilleurs résultats avec une période de calibration de 72 heures. En ce qui concerne les fenêtres d'apprentissage, nous avons constaté que l'algorithme le plus performant, à savoir la régression logistique, a obtenu ses meilleures performances avec une fenêtre d'apprentissage de 42 heures précédant la prise d'antibiotique. En d'autres termes, les prématurés du groupe LOS ont été étiquetés infectés 42 heures avant le diagnostic clinique de LOS par le clinicien. Bien que les autres algorithmes n'aient pas obtenu leurs meilleures performances avec cette fenêtre, ils ont tous atteints des performances très satisfaisantes lorsqu'ils ont été entraînés sur cette période. Par conséquent, nous avons admis que les 42 heures précédant le diagnostic clinique de LOS représentaient un bon compromis pour étiqueter les échantillons pour un algorithme d'apprentissage supervisé.

Sur le plan des performances, nous avons montré que trois des quatre algorithmes proposés obtenaient de meilleures performances lorsque les indices de graphes de visibilité étaient introduits dans l'ensemble de données. L'exception étant, une fois encore, l'algorithme SVM. L'algorithme le plus performant, la régression logistique, a atteint une surface maximale sous la courbe opérationnelle de réception (AUROC) de 87,7% lorsqu'il a été évalué dans les six heures précédant le diagnostic clinique. Cette AUROC était supérieure de 6,8% à celle que l'algorithme avait pour la même période d'évaluation lorsque les caractéristiques du graphique de visibilité étaient exclues. Une analyse du rapport de vraisemblance a confirmé que cette amélioration était statistiquement significative.

Cette étude a donc conduit à un système de décision complet associant les graphes de visibilité et les paramètres classiques liés à la variabilité cardiaque et une période de calibration initiale de 48h. La décision étant prise par régression logistique. En ce qui concerne les annotations, notre travail a permis de montrer qu'une période de 42 heures précédant le diagnostic clinique pouvait être considérée comme positive pour la période d'infection et permettre ainsi l'entraînement des algorithmes d'apprentissage supervisé.

Réseaux neuronaux récurrents pour le diagnostic précoce de l'infection chez les prématurés fondés sur la variabilité de la fréquence cardiaque

Les réseaux neuronaux récurrents (RNN) se caractérisent par des connexions récurrentes qui lui confèrent une capacité de mémoire. Par conséquent, les RNN sont particulièrement bien adaptés pour traiter les séries temporelles. Ils sont capables de détecter des modèles temporels complexes et d'identifier des dépendances. Des études antérieures ont suggéré leur utilisation pour la détection de l'infection chez l'adulte [10]. A notre connaissance, ces derniers n'ont pas été utilisés pour le diagnostic précoce de l'infection chez les prématurés fondés sur la variabilité de la fréquence cardiaque. Nous avons donc testé leur intérêt dans ce travail.

Nous avons conçu deux modèles RNN différents. Le premier utilise en entrée la série HRV temporelle brute correspondant à 1024 battements (approximativement 6 min d'enregistrement). Le second modèle exploite les paramètres usuels qui caractérisent la variabilité. Pour le second modèle, et en tenant compte de nos résultats précédents, nous avons inclus les indices du graphe de visibilité dans l'ensemble des caractéristiques. Une nouvelle fois, et en tenant compte de nos expérimentations menées au chapitre précédent, l'apprentissage des modèles a été mené sur un horizon de 42h avant le diagnostic d'infection par le clinicien. Tous les échantillons correspondant à des intervalles de temps antérieurs à cette période ont été étiquetés comme contrôles. La différence fondamentale avec l'approche du chapitre précédent est que nous n'avons pas utilisé de période de calibration. La nature même des RNN et leur capacité à prendre en compte les changements dépendant du temps ont justifié cette stratégie.

L'évaluation a été réalisée sur une population de 259 prématurés (218 dans le groupe contrôle et 41 dans le groupe LOS) répartis dans un ensemble d'apprentissage et de test. Les performances du premier modèle RNN (qui s'appuie sur les données HRV brutes) sur l'ensemble de test ont présenté une AUROC maximale de 70,7%, sur une période d'évaluation de six heures avant le diagnostic d'infection. Bien que les résultats soient médiocres, cette expérimentation met en évidence la capacité du modèle à prédire à partir de l'ensemble d'apprentissage l'infection mais induit vraisemblablement un surapprentissage. Ceci s'explique par le fait que la base de données est réduite (en regard de la taille des paramètres inclus dans le modèle). Ces constats suggèrent

de ne pas abandonner cette approche mais bien de la poursuivre avec une base de données plus importante.

En revanche, le modèle RNN qui utilisait les caractéristiques de la HRV en entrée présente des résultats intéressants. L'AUROC maximale mesurée avoisine les 90% sur l'ensemble de test pour la fenêtre d'évaluation de six heures avant le diagnostic clinique de l'infection. En outre, nous avons observé que l' AUROC était constamment supérieure à 80% pour une période de 24 heures précédant le diagnostic clinique.

Évaluation de la maturation des prématurés par un algorithme d'apprentissage ensembliste utilisant des signaux physiologiques

Dans cette deuxième partie, nous avons proposé un modèle d'apprentissage ensembliste pour l'estimation d'un âge de maturation fonctionnelle (FMA) basé sur des données physiologiques et plus particulièrement celles relevant du système cardiovasculaire. Nous avons ici supposé que le FMA, et son écart éventuel par rapport à l'âge postmenstruel (PMA) déterminé cliniquement, peut être un indicateur du niveau de maturation des prématurés.

Le modèle que nous avons proposé intègre (i) une sélection automatique des paramètres, par le biais d'une étape de filtrage qui élimine les caractéristiques qui sont très faiblement corrélées au PMA, (ii) un algorithme génétique appliqué aux caractéristiques restantes. Les caractéristiques choisies par l'algorithme génétique constituent l'entrée de l'algorithme d'apprentissage ensembliste. Ce dernier associe une régression linéaire et une régression par forêts aléatoires. Sur le plan pratique, cela signifie que deux instances de l'algorithme génétique sont appliquées : l'une pour optimiser l'ensemble des caractéristiques pour la régression linéaire, et l'autre pour optimiser l'ensemble de caractéristiques pour la régression par forêts aléatoires. Le modèle de régression linéaire (alimenté par les caractéristiques retenues pour ce modèle) est appliqué en premier et estime alors le FMA en prenant pour cible le PMA. Cette estimation est placée alors comme une caractéristique supplémentaire dans la régression par forêts aléatoires qui exploitent les autres indices choisis pour ce modèle par l'algorithme génétique. Le FMA estimé par ce second modèle est alors le résultat final de l'algorithme d'apprentissage ensembliste.

Bien que ce modèle ait été initialement conçu pour fonctionner sur les données de HRV, il a ensuite été généralisé pour traiter différents types de données disponibles dans le projet Digi-NewB. Nous avons donc testé trois instances différentes du modèle : l'une, comme déjà mentionné, sur l'ensemble des caractéristiques de la HRV, l'autre sur un ensemble de caractéristiques de la variabilité de la fréquence respiratoire (RRV), et la troisième sur un ensemble de caractéristiques dérivées des événements bradycardiques.

Les expérimentations et tests de ces modèles ont été réalisés sur une population de 50 prématurés, d'âges gestationnels différents. Dans cette première expérimentation, nous avons

retenu une population de prématurés où aucun retard neurodéveloppemental pendant la période d'évaluation n'a été observé. En d'autres termes, cette population a été jugée comme notre population contrôle. Nous avons constaté que, sur cette population, le modèle était capable d'estimer un âge FMA très proche du PMA réel des prématurés, avec une erreur absolue moyenne (MAE) de 0,93 semaine lorsqu'on utilise l'ensemble des caractéristiques HRV, et une MAE de 1,39 semaines lorsqu'on utilise les caractéristiques RRV et /ou des événements bradycardiques.

Évaluation du modèle d'apprentissage ensembliste sur une population de prématurés présentant une maturation anormale

Dans cet ultime chapitre, le modèle d'apprentissage ensembliste proposé au chapitre précédent a été exploité pour détecter des anomalies de maturation des prématurés. L'idée simple est que si le FMA estimé s'écarte de manière significative du PMA réel des prématurés alors une altération de la maturation neuro-développementale est peut-être suspectée.

Pour tester cette hypothèse, nous avons entraîné trois modèles (l'un pour la HRV, le second pour la RVV et le dernier sur les événements bradycardiques) sur une population de 40 prématurés sans anomalie de développement. Le modèle construit a été testé sur les caractéristiques correspondantes d'une population test de dix prématurés contrôles et d'une population de 54 prématurés présentant une maturation anormale. Les prématurés de cette population regroupait : (i) une population présentant des lésions neurologiques (NL) (liées en particulier soit à une hémorragie interventriculaire, soit à une leucomalacie périventriculaire); (ii) une population de dysplasie broncho-pulmonaire (DBP) ; (iii) une population de prématurés souffrant d'entérocolite nécrosante (ECN) ; (iv) et enfin une population de prématurés recouvrant deux ou plusieurs des affections précédentes.

Les performances ont été évaluées une nouvelle fois en étudiant la MAE mais aussi la corrélation avec mesures répétées entre le FMA estimé par les modèles et le PMA déterminé cliniquement. Nous avons constaté que pour les trois types de données testées, la MAE le plus faible et la corrélation la plus élevée ont été obtenues pour la population test de prématurés en bonne santé. Parmi les sous-classes de la population présentant une maturation anormale, celle qui présentait ensuite l'écart le plus faible était la DBP. En utilisant les ensembles de caractéristiques déduites de la HRV et de la RRV, la population NL a montré la MAE la plus élevée et la corrélation la plus faible avec le PMA. En exploitant les caractéristiques des bradycardies, les populations NEC puis celle avec des conditions multiples ont montré la MAE la plus grande et la plus faible corrélation.

Ces quelques résultats sont intéressants et montrent que le modèle d'estimation du FMA pourrait être utilisé pour identifier les prématurés présentant une trajectoire de maturation anormale et pour quantifier ces anomalies de croissance.

Conclusion et Travaux Futurs

En synthèse, ce travail a permis de proposer différents modèles d'apprentissage qui peuvent potentiellement être utilisés comme systèmes d'aide à la décision pour le diagnostic précoce d'infection et pour l'évaluation de la maturation des prématurés dans le contexte des unités de soins intensifs néonatales.

Pour l'infection, un modèle à base de réseaux neuronaux récurrents alimentés par des paramètres extraits de la variabilité cardiaque a permis d'atteindre une AUROC supérieure à 90%, 6h avant la décision clinique.

Pour la maturation, un modèle d'estimation de la FMA a été développé. On a montré que ce dernier peut être habilement exploité pour détecter des anomalies de croissance en USIN.

L'un des points forts de notre travail est que, pour chaque modèle que nous avons proposé, à la fois pour les objectifs d'infection et de maturation, notre souci a été de développer une preuve de concept en unités de soins intensifs capable de travailler en temps réel. Plusieurs cas d'usage sont reportés dans ce manuscrit pour illustrer notre propos. Ce résultat est majeur et montre tout le bien fondé de notre démarche en proposant un système d'aide à la décision.

Les perspectives liées à ce travail sont multiples et pourraient se concentrer sur d'autres types de données dans les modèles proposés. Dans le cas du modèle d'infection, il pourrait s'agir, par exemple, de données sur la RRV. Dans le cas du modèle d'évaluation de la maturation, on envisage à très court terme d'inclure les caractéristiques, déjà disponibles dans le projet Digi-NewB, telles que les données du mouvement, les pleurs et le sommeil.

Sur un plan plus clinique, le modèle de détection d'infection et le modèle d'évaluation de la maturation doivent maintenant être validés sur un ensemble de données plus importants, incluant de préférence des données externes au projet Digi-NewB. Enfin et comme ultime challenge, des essais cliniques long terme doivent être menés pour déterminer si les modèles proposés peuvent contribuer à la réduction de la morbidité, de la mortalité et de la durée d'hospitalisation dans les USIN.

Bibliography

[1] S. Chawanpaiboon, J. P. Vogel, A.-B. Moller, P. Lumbiganon, M. Petzold, D. Hogan, S. Landoulsi, N. Jampathong, K. Kongwattanakul, M. Laopaiboon, C. Lewis, S. Rattanakanokchai, D. N. Teng, J. Thinkhamrop, K. Watananirun, J. Zhang, W. Zhou, and A. M. Gülmezoglu, "Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis," *The Lancet Global Health*, vol. 7, no. 1, pp. e37–e46, Jan. 2019.

[2] B. M. Melnyk, N. F. Feinstein, L. Alpert-Gillis, E. Fairbanks, H. F. Crean, R. A. Sinkin, P. W. Stone, L. Small, X. Tu, and S. J. Gross, "Reducing premature infants' length of stay and improving parents' mental health outcomes with the creating opportunities for parent empowerment (COPE) neonatal intensive care unit program: A randomized, controlled trial," *Pediatrics*, vol. 118, no. 5, pp. e1414–e1427, 2006.

[3] G. J. Escobar, S. Joffe, M. N. Gardner, M. A. Armstrong, B. F. Folck, and D. M. Carpenter, "Rehospitalization in the first two weeks after discharge from the neonatal intensive care unit," *Pediatrics*, vol. 104, no. 1, pp. e2–e2, 1999.

[4] H. Khazaei, N. Mench-Bressan, C. McGregor, and J. E. Pugh, "Health informatics for neonatal intensive care units: An analytical modeling perspective." *IEEE J Transl Eng Health Med*, vol. 3, p. 3000109, 2015.

[5] I. Gur, A. Riskin, G. Markel, D. Bader, Y. Nave, B. Barzilay, F. G. Eyal, and A. Eisenkraft, "Pilot study of a new mathematical algorithm for early detection of late-onset sepsis in very low-birth-weight infants," *Am J Perinatol*, vol. 32, no. 04, pp. 321–330, 2015.

[6] L. B. Mithal, R. Yogev, H. L. Palac, D. Kaminsky, I. Gur, and K. K. Mestan, "Vital signs analysis algorithm detects inflammatory response in premature infants with late onset sepsis and necrotizing enterocolitis," *Early Human Development*, vol. 117, pp. 83–89, 2018.

[7] P. Galdi, M. Blesa, D. Q. Stoye, G. Sullivan, G. J. Lamb, A. J. Quigley, M. J. Thrippleton, M. E. Bastin, and J. P. Boardman, "Neonatal morphometric similarity mapping for predicting brain age and characterizing neuroanatomic variation associated with preterm birth," *NeuroImage: Clinical*, vol. 25, p. 102195, 2020.

[8] N. J. Stevenson, L. Oberdorfer, N. Koolen, J. M. O'Toole, T. Werther, K. Klebermass-Schrehof, and S. Vanhatalo, "Functional maturation in preterm infants measured by serial recording of cortical activity," *Scientific Reports*, vol. 7, no. 1, p. 12969, 2017.

[9] T. Nguyen Phuc Thu, A. I. Hernández, N. Costet, H. Patural, V. Pichot, G. Carrault, and A. Beuchée, "Improving methodology in heart rate variability analysis for the premature infants: Impact of the time length," *PLOS ONE*, vol. 14, no. 8, pp. 1–14, 08 2019.

[10] T. Van Steenkiste, J. Ruyssinck, L. De Baets, J. Decruyenaere, F. De Turck, F. Ongenae, and T. Dhaene, "Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks," *Artificial Intelligence in Medicine*, vol. 97, pp. 38–43, Jun. 2019.

Table of Contents

List of acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Network
ANS	Autonomic Nervous System
AUROC	Area Under the Receiver Operating Characteristics Curve
BDP	Bronchopulmonary Dysplasia
BPM	Beats Per Minute
CI	Confidence Interval
CNN	Convolutional Neural Network
DSS	Decision Support System
ECG	Electrocardiogram
EEG	Electroencephalogram
EHR	Electronic Health Records
EML	Ensemble Machine Learning
EP	Extreme Preterm
ET	Early Term
FMA	Functional Maturational Age
FT	Full Term
GA	Gestational Age
GRU	Gated Recurrent Unit
HR	Heart Rate
HRV	Heart Rate Variability
IVH	Intraventricular Hemorrhage
KNN	K-Nearest Neighbors
LogR	Logistic Regression
LOOC	Leave-one-out Cross-Validation
LOS	Late Onset Sepsis
LP	Late Preterm
LR	Linear Regression
LSTM	Long Short-Term Memory

MAE	Mean Absolute Error
MLA	Machine Learning Algorithms
MRI	Magnetic Resonance Imaging
NEC	Necrotizing Enterocolitis
NICU	Neonatal Intensive Care Unit
NL	Neurological Lesions
PAM	Population with Abnormal Maturation
PCA	Principal Component Analysis
PMA	Postmenstrual Age
PVL	Periventricular Leukomalacia
RandF	Random Forest
RFR	Random Forest Regression
RNN	Recursive Neural Network
RR	Respiration Rate
RRV	Respiration Rate Variability
SVM	Support Vector Machine
VP	Very Preterm

List of Figures

List of Tables

Introduction

Contextualizing the Problem: Incidence of Preterm Births and Associated Risks

Preterm births are defined by the World Health Organization as all births before 37 completed weeks of gestation or, equivalently, before 259 completed days from the first day of the mother's last menstrual period [1]. According to the latest statistics, based on global data from 2014, preterm births account for approximately 10.6% of live births worldwide [2]. The rate of preterm births in Europe is very similar. According to the latest Euro-Peristat report, based on data corresponding to the year 2015 from 31 European countries, the rate of preterm births among the countries in the study varied from 6% to 12% of live births [3]. In contrast, on the same Euro-Peristat report, preterm infants accounted for 73.5% of all neonatal deaths.

Besides being associated with increased morbidity and mortality [4], preterm birth is also linked with increased duration of hospital stay [5], and higher probability of hospital readmission after initial discharge [6], as well as during the first years of life [7]. Both of these factors also lead to psychological distress for the parents [8], and increased financial costs to the health care system ([9, 10, 11]).

Premature neonates are at higher risk of a plethora of short-term complications, including respiratory distress syndrome [12], early onset sepsis [13], late onset sepsis ([14, 15]), necrotizing enterocolitis ([16, 17]), intraventricular hemorrhage ([18, 19]), periventricular leukomalacia ([20, 21]), neonatal jaundice, and hypoxic-ischaemic encephalopathy [22]. Furthermore, this population also presents an increased risk of multiple long-term morbidities, including bronchopulmonary dysplasia ([23, 24]); retinopathy of prematurity [25] and other types of visual impairment [26]; hearing impairments [27]; behavioral and cognitive sequelae [28], such as attention deficit hyperactivity disorder [27], epilepsy [22], and cerebral palsy [27].

The outcome and prognosis of preterm infants is closely related with their gestational age (GA), which is defined by the American Academy of Pediatrics [29] as the time elapsed between the first day of the mother's last menstrual period and the day of delivery. In fact, all the aforementioned health risks increase with decreasing GA. Moreover, even though these conditions have a higher incidence in preterm infants, early term infants (born between 37 to 38 weeks of GA) [30] still have a higher rate of morbidity and admissions to neonatal intensive care units (NICU) than full term infants (born at 39 weeks of GA or older) [31].

The Impact of Modern Medicine in Mortality and Morbidity Rates Among Preterm Infants

Although the morbidity and mortality rates are considerably high for these vulnerable populations, that does not mean that there have not been significant improvements in the past decades in the area of neonatal medical care. Starting with the introduction of widespread use of assisted ventilation in the late 1960s and early 1970s, which saw the survival rates for preterm infants increase by more than 10% between 1968 and 1978 [32]. This was quickly followed by the introduction of the administration of antenatal corticosteroids to women at risk of preterm birth, in order to accelerate fetal lung maturation, which further helped reduce mortality rates among preterm infants, as well as some of the most common associated complications, such as respiratory distress syndrome [33]. Later, the introduction of surfactant in the late 1980s further decreased mortality among preterm infants [34].

These improvements in neonatal medicine also lead to changes in the attitude of physicians towards intensive care, which started being offered to preterm infants with very low GA which might have been considered nonviable in previous decades. This led to an increase in the number of extremely preterm infants admitted in NICUs, and a subsequent decrease in the mortality rates among these infants [35].

The combination of all these factors has lead to an increase in the survival rates of preterm infants of all GAs since the last decades of the previous century. This trend has continued into the 21st century, with Euro-Peristat recording decreases in the mortality rate of preterm infants with regard to its previous reports [3]. However, the increase in survival rates has also meant an increase in number of hospital readmission [7] and patients presenting long-term risks factors associated with preterm birth [36].

Current Challenges and Opportunities

These recent trends in morbidity and mortality in the preterm population have translated into a current need for increased research, not only on how to further reduce mortality, but also on how to improve the outcome for prematurely born patients [37]. One of the possible ways to reduce long-term morbidity is through early diagnosis of diseases that affect this population in the perinatal period.

For example, a meta-analysis of 17 studies, involving 15,331 preterm or very low birth weight infants, showed that neonates who survived sepsis in the neonatal period are at higher risk for long-term neurodevelopmental disability [38]. Consequently, early diagnosis of sepsis, leading to prompt and adequate treatment, could help reduce not only mortality but also long-term morbidity rates among preterm infants. Similarly, early diagnosis of cerebral palsy could lead to timely interventions regarding task-specific training and parental support, as well as prompt

diagnosis of comorbidities such as pain, sleep disorders, visual and hearing impairments; such early interventions have been associated with improved outcomes for these children [39].

Fortunately, the last decades have seen an increase in the research and development of early diagnosis tools in neonatal care, inspired by the popularization of big data and artificial intelligence and their application to the medical field.

Big data is a vaguely defined concept, but which in general refers to very large datasets, enclosing various features in a complex structure, which makes them hard to manage and interpret using conventional approaches ([40, 41, 42]). On the other hand, artificial intelligence has been described in the field of computer science in terms of *rational agents*, defined by Stuart Russell and Peter Norvig as computer programs that "*operate autonomously, perceive their environment, persist over a prolonged time period, and adapt to change*," with the aim to achieve "*the best outcome or, when there is uncertainty, the best expected outcome*" [43]. As such, artificial intelligence and machine learning are well suited for extracting information from big data to construct models or make predictions that can be more easily interpreted and analysed by humans [44].

In the health care sector, including neonatal medicine, large amounts of data can be acquired from electronic medical records, from periodical or continuous vital signs monitoring, and from medical imaging. This data can later be fed to artificial intelligence algorithms that can be trained to identify complex patterns in the data that lead to earlier diagnosis of a certain diseases and thus serve as a decision support system ([45, 46]).

NICUs, in particular, gather vast amounts of data, given that generally the patients' vital signs are frequently annotated, and physiological signals such as heart rate, respiration rate, and oxygen saturation are continuously monitored. In fact, it has been estimated that in a well functioning NICU, around 40 Terabytes of data are generated per bed per year [47]. The variety of information gathered in NICU can be analyzed by machine learning algorithms that can aid physicians in diagnosis and decision-making. Therefore, it is not surprising that a recent review of the use of machine learning in child and adolescent health [48] found that most of the studies in this category belonged to the field of neonatology, which has seen a persistently high level of interest in artificial intelligence and machine learning techniques throughout the last three decades.

One of the most recent projects targeted at improving neonatal health care through data gathering and the use of machine learning and artificial intelligence is the Europe-based Digi-NewB project [49].

Digi-NewB

Digi-NewB is a research project, founded by the European Union, that aims to improve health care for neonates through the development of a new generation monitoring system, with a particular focus on the detection of sepsis risk and evaluation of the maturation in premature infants.

The project was carried out by seven partners, from both public and private sectors, and coming from four European countries: France, Finland, Ireland, and Portugal. The project was conducted between March 2016 and May 2020, in collaboration with the University of Rennes 1 (France), the University of Galway (Ireland), the Institute for Systems and Computer Engineering, Technology and Science (Portugal), Tampere University (Finland), and two small and medium-sized enterprises: Syncrophi (Ireland) and Voxygen (France). It was led by the Western Network of University Hospitals in France (GCS HUGO). As part of the project, electronic health records, recordings of vital signs, video and sound data were gathered in the NICU of six university hospitals in the western region of France, namely the University Hospitals of Angers, Brest, Nantes, Poitiers, Rennes, and Tours.

The main objective of the Digi-NewB project was the development of a decision support system (DSS) that aims to assist the physician in decision-making processes. The DSS is meant to be based on a non-invasive monitoring system, and be able to aid in the early detection of sepsis, and in the quantification of the infants' cardio-respiratory and neurobehavioral maturation. Such DSS could lead to novel preventive and therapeutic strategies that could further lower mortality rates among preterm infants as well as improve their long-term outcome.

The work presented in this thesis is entirely framed in the Digi-NewB project, as it was carried out using data emerging from this project, and with the aim to help in the accomplishment of the Digi-NewB goals.

Objectives

The objectives of this dissertation are aligned with those of the Digi-NewB project, and therefore are concerned with improving sepsis risk assessment tools and proposing a method for the objective quantification of maturation of preterm infants. However, as previously mentioned, Digi-NewB collected several different types of data and is a project carried out through the collaboration of multiple partners.

Therefore, the scope of our work will be limited to applications involving cardio-respiratory data, derived from heart rate monitoring, for the achievement of the Digi-NewB goals. Specifically, the objectives of this thesis are:

— Evaluating the impact of including novel features for the characterization of heart rate variability in the performance of machine learning models for the early detection of late-

onset sepsis (LOS) in premature infants.
— Identifying optimal time windows previous to the clinical diagnosis of LOS to label as septic samples for the purpose of training supervised learning algorithms.
— Implementing a machine learning model for early LOS diagnosis based on heart rate variability data, that has the potential of being deployed as a non-invasive and real-time DSS in NICUs.
— Proposing a metric to objectively evaluate the maturational development of preterm infants, based on physiological signals, during their hospital stay.
— Proposing a machine learning model that can accurately assess the maturation of preterm neonates based on the aforementioned metric, and which can detect disruptions in the normal maturation pattern of the infants.

Outline

The remainder of this dissertation is organized as follows:

— In Chapter 1 we review some background concepts that will be used throughout the rest of the dissertation, such as artificial intelligence and machine learning. Then, we present the state of the art in regards to the use of machine learning techniques for the early diagnosis of sepsis in premature infants and for the evaluation of the neonates' maturation. Finally, we describe in more detail the Digi-NewB project, in which all subsequent studies presented in the thesis are framed.
— In Chapter 2 we study the impact of different training windows and the inclusion of the visibility graph indexes as additional features for the characterization of the heart rate variability, and their impact on the performance of four different types of machine learning algorithms. This study was done with a population of 49 premature infants.
— In Chapter 3 we proposed a recurrent neural network model, which uses heart rate variability data, to diagnose LOS in preterm infants. This method was developed with a population of 259 infants, and for its design we took into consideration the findings presented in the previous chapter.
— In Chapter 4 we design, develop, and test an ensemble machine learning model on a population of 50 healthy infants, for the purpose of evaluating their maturation through the estimation of their maturational age. The model proposed in this chapter was tested using heart rate variability, respiration rate variability, and bradycardia data.
— In Chapter 5 we validate the model presented in Chapter 4 on both healthy infants, based on the same population used in said chapter, and on a population of preterm infants who were diagnosed in the postnatal period with medical conditions that have been documented to have negative impact in the neurodevelopmental outcome.

— In the Conclusion we offer some final remarks about the outcome of the research presented in this dissertation, and present a summary of our findings, regarding both the LOS and maturation objectives, as well as the strengths and limitations of our study. Finally, we present some insights regarding possible future directions to continue this line of research.

Bibliography

[1] "WHO: recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. modifications recommended by FIGO as amended october 14, 1976." *Acta Obstet Gynecol Scand*, vol. 56, no. 3, pp. 247–253, 1977.

[2] S. Chawanpaiboon, J. P. Vogel, A.-B. Moller, P. Lumbiganon, M. Petzold, D. Hogan, S. Landoulsi, N. Jampathong, K. Kongwattanakul, M. Laopaiboon, C. Lewis, S. Rattanakanokchai, D. N. Teng, J. Thinkhamrop, K. Watananirun, J. Zhang, W. Zhou, and A. M. Gülmezoglu, "Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis," *The Lancet Global Health*, vol. 7, no. 1, pp. e37–e46, Jan. 2019.

[3] Euro-Peristat Project. European Perinatal Health Report. (2018) Core indicators of the health and care of pregnant women and babies in europe in 2015. [Online]. Available: www.europeristat.com

[4] Y. Dong and J.-L. Yu, "An overview of morbidity, mortality and long-term outcome of late preterm birth," *World Journal of Pediatrics*, vol. 7, no. 3, p. 199, 2011.

[5] B. M. Melnyk, N. F. Feinstein, L. Alpert-Gillis, E. Fairbanks, H. F. Crean, R. A. Sinkin, P. W. Stone, L. Small, X. Tu, and S. J. Gross, "Reducing premature infants' length of stay and improving parents' mental health outcomes with the creating opportunities for parent empowerment (COPE) neonatal intensive care unit program: A randomized, controlled trial," *Pediatrics*, vol. 118, no. 5, pp. e1414–e1427, 2006.

[6] G. J. Escobar, S. Joffe, M. N. Gardner, M. A. Armstrong, B. F. Folck, and D. M. Carpenter, "Rehospitalization in the first two weeks after discharge from the neonatal intensive care unit," *Pediatrics*, vol. 104, no. 1, pp. e2–e2, 1999.

[7] L. W. Doyle, G. Ford, and N. Davis, "Health and hospitalisations after discharge in extremely low birth weight infants," *Seminars in Neonatology*, vol. 8, no. 2, pp. 137–145, 2003.

[8] L. T. Singer, A. Salvator, S. Guo, M. Collin, L. Lilien, and J. Baley, "Maternal psychological distress and parenting stress after the birth of a very low-birth-weight infant." *JAMA*, vol. 281, no. 9, pp. 799–805, Mar 1999.

[9] S. Petrou, G. Abangma, S. Johnson, D. Wolke, and N. Marlow, "Costs and health utilities associated with extremely preterm birth: evidence from the epicure study." *Value Health*, vol. 12, no. 8, pp. 1124–1134, Nov-Dec 2009.

[10] A. Bérard, M. Le Tiec, and M. A. De Vera, "Study of the costs and morbidities of late-preterm birth," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 97, no. 5, pp. F329–F334, 2012. [Online]. Available: https://fn.bmj.com/content/97/5/F329

[11] S. Petrou, H. H. Yiu, and J. Kwon, "Economic consequences of preterm birth: a systematic review of the recent literature (2009-2017)." *Arch Dis Child*, vol. 104, no. 5, pp. 456–465, May 2019.

[12] D. G. Sweet, V. Carnielli, G. Greisen, M. Hallman, E. Ozek, R. Plavka, O. D. Saugstad, U. Simeoni, C. P. Speer, M. Vento, and H. L. Halliday, "European consensus guidelines on the management of neonatal respiratory distress syndrome in preterm infants - 2013 update," *Neonatology*, vol. 103, no. 4, pp. 353–368, 2013.

[13] S. Mukhopadhyay and K. M. Puopolo, "Risk assessment in neonatal early onset sepsis," *Seminars in Perinatology*, vol. 36, no. 6, pp. 408–415, 2012.

[14] B. J. Stoll, N. Hansen, A. A. Fanaroff, L. L. Wright, W. A. Carlo, R. A. Ehrenkranz, J. A. Lemons, E. F. Donovan, J. E. T. Ann R. Stark, W. Oh., C. R. Bauer, S. B. Korones, S. Shankaran, A. R. Laptook, D. K. Stevenson, L.-A. Papile, and W. K. Poole, "Late-Onset Sepsis in Very Low Birth Weight Neonates: The Experience of the NICHD Neonatal Research Network," *Pediatrics*, vol. 110, no. 2, pp. 285–291, Aug. 2002.

[15] Y. Dong and C. P. Speer, "Late-onset neonatal sepsis: recent developments," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 100, no. 3, pp. F257–F263, 2015.

[16] A. M. Thompson and M. J. Bizzarro, "Necrotizing enterocolitis in newborns," *Drugs*, vol. 68, no. 9, pp. 1227–1238, 2008.

[17] L. Berman and R. L. Moss, "Necrotizing enterocolitis: An update," *Seminars in Fetal and Neonatal Medicine*, vol. 16, no. 3, pp. 145–150, 2011.

[18] H. J. McCrea and L. R. Ment, "The diagnosis, management, and postnatal prevention of intraventricular hemorrhage in the preterm neonate," *Clinics in Perinatology*, vol. 35, no. 4, pp. 777–792, 2008.

[19] J. S. von Lindern, T. van den Bruele, E. Lopriore, and F. J. Walther, "Thrombocytopenia in neonates and the risk of intraventricular hemorrhage: a retrospective cohort study," *BMC Pediatrics*, vol. 11, no. 1, p. 16, 2011.

[20] J. J. Volpe, "Neurobiology of periventricular leukomalacia in the premature infant," *Pediatric Research*, vol. 50, no. 5, pp. 553–562, 2001.

[21] E. Hatzidaki, E. Giahnakis, S. Maraka, E. Korakaki, A. Manoura, E. Saitakis, I. Papamastoraki, K.-M. Margari, and C. Giannakopoulou, "Risk factors for periventricular leukomalacia," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 88, no. 1, pp. 110–115, 2009.

[22] M. Platt, "Outcomes in preterm infants," *Public Health*, vol. 128, no. 5, pp. 399–403, 2014.

[23] J. P. Kinsella, A. Greenough, and S. H. Abman, "Bronchopulmonary dysplasia," *The Lancet*, vol. 367, no. 9520, pp. 1421–1431, 2006.

[24] C. L. Day and R. M. Ryan, "Bronchopulmonary dysplasia: new becomes old again!" *Pediatric Research*, vol. 81, no. 1, pp. 210–213, 2017.

[25] H. Blencowe, J. E. Lawn, T. Vazquez, A. Fielder, and C. Gilbert, "Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010," *Pediatric Research*, vol. 74, no. 1, pp. 35–49, 2013.

[26] A. R. O'Connor, C. M. Wilson, and A. R. Fielder, "Ophthalmological problems associated with preterm birth," *Eye*, vol. 21, no. 10, pp. 1254–1260, 2007.

[27] S. Saigal and L. W. Doyle, "An overview of mortality and sequelae of preterm birth from infancy to adulthood," *The Lancet*, vol. 371, no. 9608, pp. 261–269, 2008.

[28] D. W. A. Milligan, "Outcomes of children born very preterm in europe," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 95, no. 4, pp. F234–F240, 2010.

[29] "Age terminology during the perinatal period," *Pediatrics*, vol. 114, no. 5, pp. 1362–1364, 2004.

[30] "Committee opinion no 579: Definition of term pregnancy," *Obstetrics & Gynecology*, vol. 122, no. 5, 2013.

[31] M. Delnord and J. Zeitlin, "Epidemiology of late preterm and early term births – an international perspective," *Seminars in Fetal and Neonatal Medicine*, vol. 24, no. 1, pp. 3–10, 2019.

[32] L. Mutch, M. Newdick, A. Lodwick, and l. Chalmers, "Secular changes in rehospitalization of very low birth weight infants," *Pediatrics*, vol. 78, no. 1, pp. 164–171, 1986.

[33] S. F. P. R. McGoldrick, E and S. Dalziel, "Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth," *Cochrane Database of Systematic Reviews*, no. 12, 2020.

[34] R. M. Schwartz, A. M. Luby, J. W. Scanlon, and R. J. Kellogg, "Effect of surfactant on morbidity, mortality, and resource use in newborn infants weighing 500 to 1500 g," *New England Journal of Medicine*, vol. 330, no. 21, pp. 1476–1480, 1994.

[35] E. Gultom, L. Doyle, P. Davis, A. Dharmalingam, and E. Bowman, "Changes over time in attitudes to treatment and survival rate for extremely preterm infants (23–27 weeks' gestational age)," *Australian and New Zealand Journal of Obstetrics and Gynaecology*, vol. 37, no. 1, pp. 56–58, 1997.

[36] J. M. Lorenz, D. E. Wooliever, J. R. Jetton, and N. Paneth, "A quantitative review of mortality and developmental disability in extremely premature newborns." *Arch Pediatr Adolesc Med*, vol. 152, no. 5, pp. 425–435, May 1998.

[37] S. W. Wen, G. Smith, Q. Yang, and M. Walker, "Epidemiology of preterm birth and neonatal outcome," *Seminars in Fetal and Neonatal Medicine*, vol. 9, no. 6, pp. 429–435, 2004.

[38] B. Alshaikh, K. Yusuf, and R. Sauve, "Neurodevelopmental outcomes of very low birth weight infants with neonatal sepsis: systematic review and meta-analysis," *Journal of Perinatology*, vol. 33, no. 7, pp. 558–564, 2013.

[39] A. J. Spittle, C. Morgan, J. E. Olsen, I. Novak, and J. L. Y. Cheong, "Early diagnosis and treatment of cerebral palsy in children with a history of preterm birth," *Clinics in Perinatology*, vol. 45, no. 3, pp. 409–420, 2021/04/15 2018.

[40] S. Madden, "From databases to big data," *IEEE Internet Computing*, vol. 16, no. 3, pp. 4–6, 2012.

[41] S. Sagiroglu and D. Sinanc, "Big data: A review," in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 2013, pp. 42–47.

[42] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[43] S. Russell and P. Norvig, *Artificial intelligence : a modern approach.* Third edition. Upper Saddle River, N.J. : Prentice Hall, ©2010, [2010].

[44] P. Ongsulee, "Artificial intelligence, machine learning and deep learning," in *2017 15th International Conference on ICT and Knowledge Engineering (ICT KE)*, 2017, pp. 1–6.

[45] S. A. Bini, "Artificial intelligence, machine learning, deep learning, and cognitive computing: What do these terms mean and how will they impact health care?" *The Journal of Arthroplasty*, vol. 33, no. 8, pp. 2358–2361, 2018.

[46] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems." *J Glob Health*, vol. 8, no. 2, p. 020303, Dec 2018.

[47] H. Khazaei, N. Mench-Bressan, C. McGregor, and J. E. Pugh, "Health informatics for neonatal intensive care units: An analytical modeling perspective." *IEEE J Transl Eng Health Med*, vol. 3, p. 3000109, 2015.

[48] Z. Hoodbhoy, S. Masroor Jeelani, A. Aziz, M. I. Habib, B. Iqbal, W. Akmal, K. Siddiqui, B. Hasan, M. Leeflang, and J. K. Das, "Machine learning for child and adolescent health: A systematic review," *Pediatrics*, vol. 147, no. 1, 2021.

[49] Digi-NewB. (2020) Digi-newb a new generation monitoring system in neonatology. [Online]. Available: http://www.digi-newb.eu/

CHAPTER 1

Smart Care in Neonatalogy and the Digi-NewB Proposal

In this chapter we present the concepts of artificial intelligence and machine learning, which will serve as a background to our work. We also offer a brief description of the machine learning algorithms that will be used throughout this dissertation. Then, we discuss the state of the art regarding machine learning approaches in the field of neonatology, focusing specifically in the two areas of interest for the present work: early diagnosis of late onset sepsis, and evaluation of the maturation of prematurely born infants. Finally, we discuss the Digi-NewB project, in which this work is framed, and how its proposal could improve the advances that have already been achieved in this field, which could help reduce mortality rates, length of hospitalization, and long-term risk factors for preterm infants.

1.1 Artificial Intelligence and Machine Learning

Artificial intelligence (AI) was broadly defined by computer scientist and one of the founders of the field, Professor John McCarthy, as follows [1]:

> The science and engineering of making intelligent machines, especially intelligent computer programs.

This definition leads to many questions about what is intelligence and how to determine if a machine or a computer program is intelligent. The answer to these questions usually comes in the form of comparison to human intelligence or human performance, which might also vary considerably. Stuart Russell and Peter Norvig offer a more precise definition of artificial intelligence, by referencing to computers that are capable of operating autonomously, perceiving their environment and being able to adapt to it, and with the goal to achieve the best expected outcome at a certain task [2].

Machine learning is considered to be a branch of AI [3], which has been defined as computer algorithms that can improve automatically, or in other words *learn*, through experience [4]. This is in fact the key difference between AI and machine learning: while the only requirement of AI is that it is or seems smart, regardless of how this is achieved, machine learning requires that it is achieved through learning from experience [5]. However, it is common to find the two terms being used interchangeably, as many advancements in the field of AI have happened specifically in the subset of machine learning.

The history of AI and machine learning dates back to the 1950s. However, it became more popular and widely used in the last decades [2], with the improvement of computers and the widespread use of Internet, which led to the production and storage of large amounts of data, from which machine learning algorithms (MLA) can learn. Currently, there is a large variety of MLAs, each of which can be applied to a multitude of seemingly very different problems. There are also many ways in which to classify MLAs.

One of the most common classifications is based on how they learn. In this sense, there are four main categories [2]:

Supervised Learning. In supervised learning the algorithm learns by observing labeled data. This means that it must receive input examples as well as the output associated with them, and then learn the functions that map each input to its associated output.

Reinforcement Learning. In this type of learning the algorithm receives rewards or punishments based on the results it gives. These rewards and punishments are commonly presented in the form of cost functions, which associate wrong results to high costs and correct results to lower costs. The algorithm then must be able to decide which of its actions prior to the reinforcement were most responsible for the outcome and improve accordingly.

Unsupervised learning. In this case only the input is given, without an associated output. The MLA must learn by finding patterns in this unlabeled data. The most common application of unsupervised learning is clustering, in which the algorithm learns to group the input examples into clusters.

Semi-supervised Learning. For semi-supervised learning the MLA must learn from a large dataset of unlabeled examples, with some labeled examples in it.

Another way to classify MLAs is based on the type of output they are trained to produce. In this regard they are mainly classified in three types:

Classification Algorithms. The output of these algorithms is a finite set of values. It might be binary or Boolean classification, if there are only two classes of possible outputs, or multiclass, if there are more than two categories of possible outputs [2].

Regression Algorithms. This type of MLAs produces continuous numeric values as output. These algorithms learn to approximate as best as possible the true value of the output based on conditional expectations given by the input [2].

Clustering Algorithms. These algorithms find patterns in order to produce as output a partition of the data into sub-groups. The main difference with classification algorithms is that clustering is associated with unsupervised learning, so the number of clusters or sub-groups in which it can possibly partition the data is not known a priori, unlike in the classification algorithms for which the number of possible classes is known [6].

Finally, MLAs can also be classified into different models based on the type of algorithm they employ. There is plethora of types of machine learning algorithms or models that are frequently used for a large variety of problems. In the next paragraphs we give an overview of only those that have been employed in this work or that have most often been used in recent projects aimed to improve neonatal care.

1.1.1 K-Nearest Neighbors

K-nearest neighbors models (KNN) are a type of supervised learning algorithm, and they are among the simplest types of machine learning. They can be applied both to classification, either multiclass and binary, and to regression problems, and they work similarly to lookup tables. Given a query x_q, the algorithm finds the k examples (or neighbors) that are nearest to x_q[2].

For classification problems, KNN assigns the class label of the majority of the k nearest neighbors in the data space [7]. To avoid ties, an odd number is usually chosen as the value of k. For regression problems, the mean or median of the k nearest neighbors is returned, or a linear regression can be fitted on the neighbors [2].

1.1.2 Linear Models

Linear models are among the simplest and oldest machine learning approaches, as they originate from statistical modeling [8]. Linear models, as their name suggests, are based on linear functions, and they can be used for regression problems as well as for binary classification problems. In the first case they are referred to as *linear regression models*, and in the later as *logistic regression*.

Linear Regression

Linear regression (LR) models output a continuous value. In the case of a multivariate problem, where the input has more than one variable attribute or feature, the input can be defined as the vector X, given by [x_1, x_2, ... x_j], with j as the number of variables features in the data, and by defining W as the vector [w_0, w_1, ... w_j], $h_W(X)$ can be defined as:

$$h_W(X) = w_0 + w_1x_1 + w_2x_2 + ... + w_jx_j \tag{1.1}$$

where w_0, w_1, ... w_j are the weights to be learned by the model, and the task of the linear model is to find the function $h_W(X)$ that best fits the data [2].

A loss function that compares the value of $h_W(X)$ to the true output y (given in the labeled examples) is used by the algorithm to find the best fit to the data. A common choice is the L2 loss function, given by:

$$\text{Loss}(\text{h}_W(\text{X})) = \sum_{n=1}^{N} (y_n - h_W(X_n))^2 \tag{1.2}$$

where N is the total number of examples in the training data, and y_n is the true output for a given example n.

Thus, the algorithm learns the weights that best fit the data by minimizing the loss function. In this sense, linear models use supervised learning, because they need labeled data to learn from, but they also use reinforcement learning, as incorrect results during training yield a high cost or loss, which the algorithm uses for learning.

Logistic Regression

Logistic regression (LogR) models are used for binary regression, and are very similar to the LR models. The main difference is on how the function $h_W(X)$ is calculated. For logistic regression $h_W(X)$ is given by the sigmoid function:

$$h_W(X) = \frac{1}{1 + e^{-X \cdot W}} \tag{1.3}$$

Thus, the predictions are values between zero and one, which can be interpreted as the probability of the sample belonging to the class labeled *1*.

As with LR, these models learn the best values for the weights W by minimizing a cost function, which can also be the *L2* function given in Equation 1.2.

1.1.3 Random Forest

Random forests are a type of supervised machine learning algorithm, based on the idea of decision trees. Decision trees are functions that take a vector of attributes or features as input, and returns a single output value [2]. They reach the output by performing tests at each node based on one or more input features, and then each node is branched into the possible values of those features. A tree might have several layers of nodes and branches, until it reaches the leaf nodes, which specify the output to be returned. Both the input and output of the decision trees can be discrete or continuous.

Random forests are built by using multiple decision trees, where each decision tree only takes as input a randomly chosen subset of the input data. Thus, to generate the n_{th} tree, a random vector θ_n is generated, independent of the past random vectors but with the same distribution

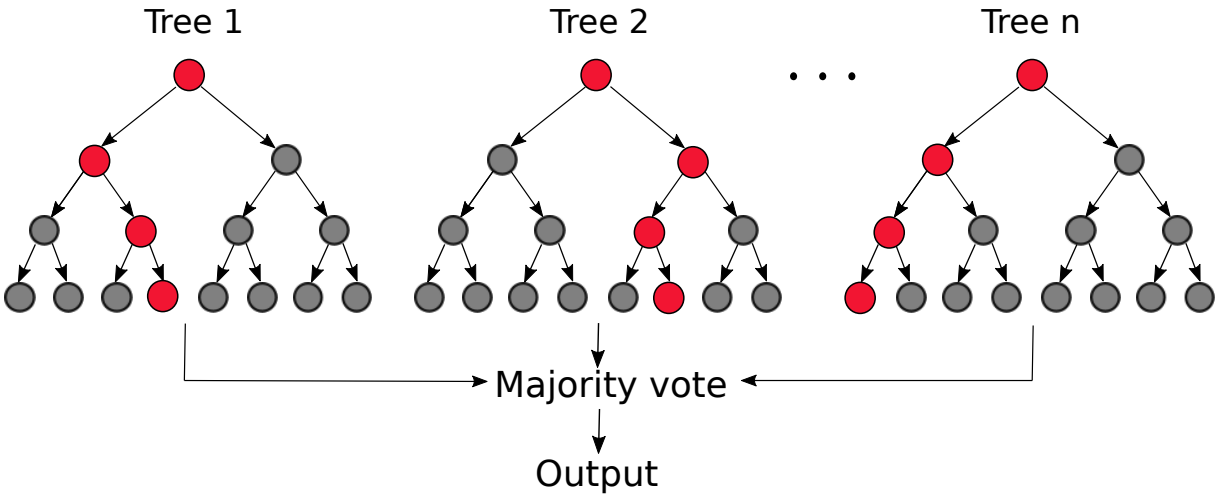


Figure 1.1 – Diagram of a simple random forest classifier

[9]; these vectors are generated as either a subset of the total examples in the input data, or as a subset of the total features associated with each example, or a combination of both. This technique helps the random forest avoid overfitting to the training data. The final output of the random forest is usually given by a majority vote of the all decision trees in the forest [10].

Same as for the decisions trees which construct it, random forest's input and output can be discrete or continuous, hence they can be applied to either classification or regression problems. Random forest models applied to classification problems are commonly referred to simply as *Random Forest* (RF). When they are applied to regression problems, they are referred to as *Random Forest Regression* (RFR).

In Figure 1.1 we show an example of a simple random forest, with n decision trees. The nodes are represented by the circles, and the branches by the arrows. We observe that each tree has four layers of nodes. In this case, each node gives only two branches, except for the nodes in the bottom row, which are the leaf nodes which represent possible final outcomes. Highlighted in red are the nodes that represent the trees' decision at each layer. Finally, the decision from each tree participates in a majority vote, the result of which is the final output.

1.1.4 Support Vector Machines

Support vector machine (SVM) models are a popular type of supervised learning algorithms. They are most commonly used for either binary or multiclass classification problems [11], but can also be extended for use in regression problems [12].

SVMs work by creating a linear separating hyperplane, given as [2]:

$$0 = W \cdot X + b \tag{1.4}$$

where X is the input, W is the vectors of the weights to be learned by the model, and has as many elements as features there are in the dataset, and b is the real-valued intercept. This is similar to the linear regression equation. However, instead of minimizing the expected empirical loss on the training data, SVMs attempt to minimize expected generalization loss, under the assumption that any future samples will be drawn from the same distribution as the training samples. This is achieved by choosing the hyperplane that not only separates the training samples, but which is also furthest from all of these samples; this hyperplane is referred to as the maximum margin separator [2].

When the data is not linearly separable by a hyperplane in the original input space, SVMs map the data into a higher dimensional feature space by means of nonlinear mapping functions [13], which are referred to as kernel functions [14]. In the feature space the data can be separated by a hyperplane given by Equation 1.4, which is not linear in the original input space [2].

1.1.5 Genetic Algorithms

Genetic algorithms are population-based stochastic machine learning algorithm, inspired by Charles Darwin's theory of evolution [15]. They fall under the category of reinforcement learning and are typically used for optimization problems.

Genetic algorithms start with a set of randomly generated states, referred to as the population, where each state is an individual or chromosome. Each chromosome corresponds to a possible solution and is composed by a finite number of parameters, or genes. Each individual is then evaluated using a fitness function and rated according to their fitness score.

The next generation is built by randomly choosing individuals from the current generation; this can be done through different functions, but always giving the fitter individuals a higher probability of being chosen. At this points, pairs of the chosen individuals (parent individuals) are combined by swapping genes with each other, thus producing two new individuals (offspring individuals); this step is called crossover or recombination. Finally, the genetic algorithms also use the idea of mutation, by randomly selecting one or more genes in each offspring individual to have its value changed. The mutation rate should be set low, or otherwise the genetic algorithm becomes a random research, but it is an important step in order to maintain the diversity of the population, which helps to avoid local solutions ([2, 15]).

1.1.6 Artificial Neural Networks

Artificial neural networks (ANN) are inspired in biological neural networks. They fall under the category of supervised learning, and both their input and output can be discrete or continuous, which makes them well suited for either regression or classification problems.

Similarly to biological neural networks, ANNs are built by combining multiple nodes or units, which are the equivalent to neurons in biological neural networks. And, like neurons,

these receive inputs from several other nodes, and send their output to several other units. The units themselves are very simple, as they take the input and combine it in a weighted sum, similar to the one presented in Equation 1.1 for LR models, and then applys to the result an activation function ([2, 16]). The activation function is a threshold function, for which common choices are the sigmoid function, hyperbolic tangent [17], and rectified linear units (ReLU) [18].

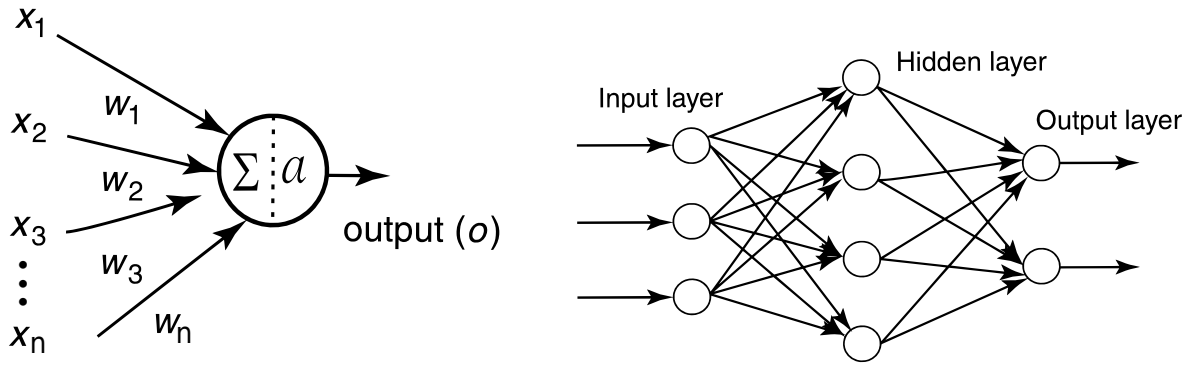
This is exemplified in Figure 1.2a, where we show a diagram of a typical unit or neuron in an ANN. In this case, the vector of x_1 through x_n is the input of the unit, each of which is multiplied by its corresponding weight (w) before being added. The result of this weighted sum is then passed to the activation function (a), which yields the final output of the unit. The values of the weights are the parameters which the algorithm has to adjust through learning from the training data.

The complexity and high performance of ANN is then given by combining multiple units. The typical architecture for this is through layers. The first layer of the network is called input layer, and receives as input the features of the dataset, and the last layer is the output layer. Between these two layers there might be multiple hidden layers, with multiples units in each layer. Every unit in a given hidden layer receives as inputs the results of the activation functions of every unit in the previous layer, and sends the output of its own activation function to every unit in the next layer.

Figure 1.2b shows a very simple example of this architecture. In this case, the input layer has only three units, and is followed by a single hidden later with four units in it. This type of layers in which every unit receives connections from every unit in the layer before it, and sends its output to every unit in the next layer are also referred to as fully connected layers. The final layer is the output layer, which in this diagram has two units. However, the number of units in each layer, as well as the number of hidden layers, might vary great. Usually, the more layers and units per layer the ANN has, the more powerful it is, but also the more computationally expensive it is to train and the most likely it is to overfit to the training data. Hence, very complex networks require more data to perform at their best capacity.

Furthermore, there are typically two distinct categories of ANN based on their architecture: *feed-forward networks*, and *recurrent neural networks* (RNN). Feed-forward networks are characterized by having connections in only one direction; an example of this type of architecture is the diagram presented in Figure 1.2b. Within this category fall the convolutional neural networks (CNN) [19], which are a type of ANN specially well suited for the analysis of images and which have been widely used to this end in recent years.

On the other hand, RNNs are characterized by having feedback connections. This implies that in this type of architecture, the response of the network to a given input depends on its initial state, which may depend on previous inputs. As a result, this type of ANN have short-term memory and, therefore, are well suited to analyse time series and sequential data [20].



(a) Diagram of a single unit in a neural network.

(b) Diagram of a simple ANN architecture.

Figure 1.2 – Diagram of typical units and architecture for ANNs. Adapted from [21]

1.2 Artificial Intelligence in Neonatal Medicine: State of the Art

Neonatology has been among the medical specialties that have witnessed most advancements in the application of artificial intelligence and machine learning to design diagnostic and prognostic tools to help clinicians better assess their patients [22]. This fact is likely the result of a combination of factors.

On one hand, patients admitted to neonatal intensive care units (NICU) have elevated rates of morbidity and mortality, and are at high risk of infections, respiratory problems, neurological problems, among others. Neonatology could benefit from machine learning approaches to help in the diagnosis of any of these conditions, which makes it an area full of opportunities for the integration of AI.

On the other hand, NICUs produce large amounts of data, coming from the continuous monitoring of the patients' vital signs, such as heart rate, respiration rate, and oxygen saturation. Additional data regarding clinical signs is also meticulously measured regularly and frequently by nurses and physicians. Further tests such as blood cultures, electroencephalograms (EEG), magnetic resonance imaging (MRI), and laboratory tests are also often done to NICU patients to better assess their situation. This abundance of data allows for AI and machine learning applications to flourish in this field.

As a consequence, machine learning approaches in neonatology vary a lot on the type of data they take as input, with some studies focusing exclusively on one type of data (for example, heart rate, or MRI images), and other studies combining several types of data. Likewise, there is also a lot of variety in the objective of the machine learning approaches being proposed, ranging from the prediction of clinically relevant hyperbilirubinemia [23], to the identification of preterm infants at risk of presenting language impairment [24].

Therefore, in the following sections we present an overview of the most recent and relevant studies on the application of MLAs in neonatal medicine focusing on the two targets which

concern this dissertation: the early detection of late onset sepsis (LOS), and the evaluation of the maturation of the premature infants.

1.2.1 Detection of Late Onset Sepsis

A common approach in current literature for LOS detection through MLAs includes the use of data available in electronic health records (EHR). One of the studies with this approach was proposed by Mani et al. [25], in which they used antibiotic, microbiology, laboratory test, and clinical data from the EHRs of 299 infants from a single hospital; while the antibiotic and microbiology data was used only for generating the labels for the dataset, laboratory and clinical data were used as the input to train the MLAs. Regarding the type of algorithm, this study tested nine different types of MLAs, including SVM, KNN, LogR, RF, decision trees, and a variety of Bayesian models [2]. The best performance was obtained by one of the Bayesian models (Naïve Bayes), with an area under the receiver operator characteristics curve (AUROC) ([26, 27, 28]) of 78%.

A similar approach was taken in a study by Masino et al. [29], in which EHR data from 618 infants was used to evaluate the performance of several MLAs, which included KNN, SVM, LogR, RF, Naïve Bayes, AdaBoost [30], and gradient boost [31]. This study did not distinguish between early and late onset sepsis, however it did differentiate two sepsis groups: one included only the infants diagnosed with sepsis who had a positive blood culture, and the other included all infants diagnosed with sepsis, regardless of whether they had a positive blood culture or not. They achieved the best results with the second population, achieving an AUROC of 89% for predicting sepsis four hours before clinical recognition.

While these approaches have good results, and there is clearly clinical and predictive value in data from laboratory test results, including this type of data has the disadvantage of requiring invasive test, which cannot be performed so frequently and also need time to turn back result. This makes these approaches invasive, and unfit to give a real-time, continuous prediction of sepsis. However, excluding laboratory tests results might come at the cost of loosing valuable information.

Nonetheless, there have been several studies dedicated to the study of machine learning approaches for LOS detection using only vital and clinical signs. One of the most prominent proposals is RALIS, an algorithm developed by Integralis Ltd. (Israel) [32]. While the MLAs it employs are not specified, this approach relies only on routinely measured signs: heart rate, respiration rate, episodes of bradycardia, oxygen saturation, body temperature, and weight.

For the RALIS approach, all the aforementioned vital signs should be measured every two hours, except the weight which should be measured every 24 hours. These measurements must be recorded into the RALIS system by the medical personal, in parallel to the routine medical documentation. This system also needs a 72-hour training period to produce a patient-specific

calibration [33]. This approach was initially tested on a population of 118 infants from two different NICUs, achieving an AUROC of 82% for detecting LOS on the same day of clinical diagnosis. Furthermore, it showed it could detect sepsis as early as three days before clinical diagnosis [33]. RALIS was later validated by Mithal et al. [34], with a cohort of 155 infants from a different hospital. They reported an AUROC of 89.9%, and that the RALIS detection occurred, on average, 33 hours before clinical suspicion.

This approach shows that MLAs can have a good performance detecting LOS even if laboratory tests are excluded from the input. However, it has the disadvantage of needing inputs to be given by the medical personnel every two hours, as it requires manual annotation in parallel to the usual medical record. This makes it prone to human error, incurs in more work for medical personnel, and because of the time delay of two hours between inputs, it cannot produce a real-time, continuous detection.

An approach that overcomes both these disadvantages is the HeRO monitoring system [35]. This system is based on the study by Griffin et al. [36], which proposes the evaluation of the risk of sepsis through logistic regression and relying exclusively on features derived from the heart rate. Specifically, they used three features derived from the time series formed by the inter-beat intervals to characterize heart rate variability (HRV), which were the standard deviation, sample entropy, and sample asymmetry. The method was developed on a population of 316 infants from the NICU of one hospital, and validated on 317 infants from the NICU of a different hospital. The HeRO system has the advantage of being able to receive as input electrocardiographic data collected from existing NICU monitors, without need for additional sensors to be in direct contact with the infants, and produces a score which is calculated in real-time and updated hourly.

The original study [36] reports an AUROC of 75%, which is lower than the AUROC reported by other studies we have discussed. However, a randomized clinical trial conducted on 3003 infants from NICUs of nine different hospitals across the United States of America, reported that the mortality rate of the HeRO display group was 22% lower than that of the nondisplay group, with no statistically significant differences in the other outcomes [37].

Although the HeRO results are impressive, it is noteworthy that it only uses three features to characterize the HRV. There are currently multiple features that are accepted in the literature for the characterization of HRV [38], including time-domain and frequency-domain features, as well as non-linear measurements. In addition, features derived from visibility graphs constructed from the inter-beat time series have recently been suggested as candidates to improve HRV analysis for neonatal sepsis detection, after promising results in animal experimentation [39]. These facts raise the question of whether a similar approach, based solely on heart rate data but including more features, could achieve even better results.

Another observation regarding the studies discussed in this section is that they use more

traditional MLAs. Therefore, it is worth considering if better results could be obtained in the detection of LOS by implementing approaches based on more complex techniques, such as ANNs.

1.2.2 Evaluation of the Maturation

Several studies have aimed at predicting the outcome of prematurely born infants based on their MRI data. One of such studies, proposed by Saha et al. [40], used diffusion MRI acquired when the infants had between 29 and 35 weeks of postmenstural age (PMA) as input data to predict abnormal motor outcome. The infants included in the study underwent clinical evaluation at two years of corrected age [41] to determine their true motor outcome. With this data, the researchers trained and tested a CNN, and achieved an AUROC of 72% on the identification of infants who had an abnormal motor outcome.

A similar approach was used in several studies by a group of researchers based in Stanford, California, to predict the outcome preterm infants would have at 18 months to two years of age, on a variety of conditions. Similarly to the study by Saha et al., these studies used diffusion MRI acquired at near term PMA, and clinical evaluations done at 18 months to two years of age to construct their datasets. The prediction method was different, however, as this group of studies favored linear models. Nonetheless, in the first of the studies they were able to predict cognitive impairment with an AUROC of 100%, and motor impairment with an AUROC of 91.2% [42]. In a second study, they were able to classify correctly the preterm infants at high risk of impaired gait velocity, reporting 93% sensitivity and 79% specificity [43], as well as predict specific characteristics of the gait also with very high sensitivity and specificity. In their most recently published studied, they successfully identified the infants at high risk for different types of language impairments, also achieving high sensitivity and specificity [24].

These studies suggest that long-term prognosis could be achieved through machine learning techniques, based on MRI data acquired during the hospitalization period. This could allow for early intervention, leading to better management and possible better outcomes for preterm infants who develop these conditions.

Another approach, also based on MRI data, was proposed by Smyser et al. [44]. In their study, they acquired resting state functional MRI data from 50 preterm infants when they were at term-equivalent age, and from 50 term-born infants within their first week of life. Then they used SVM regression to estimate the birth gestational age (GA) of the infants. The mean difference between the true GA and the GA estimated by this method was four weeks, and true and estimated GAs showed to be highly correlated, with a p-value of 0.98.

A more recent study, by Galdi et al. [45], uses a very similar approach, by taking MRI data from 59 preterm infants and 46 term infants who were scanned between 38 and 45 weeks of PMA. The images were processed and fed to a linear regression algorithm with the target to predict, in this case, the PMA at the time of the images' acquisition, and SVM to classify the

infants in preterm and full-term populations. This method was able to estimate the PMA with a mean absolute error (MAE) of 0.70 ± 0.56 weeks, and was able to correctly classify the infants as preterm or full-term with 92% accuracy.

The results obtained by Smyser et al. and by Galdi et al. suggest that the machine learning approaches are capable of detecting disruptions in the brain maturation of preterm infants, and characterize typical maturation based on MRI data. However, one shortcoming of this approach is that it evaluates the preterm infants only at term-equivalent age, (meaning, at 38 weeks of PMA or older), which makes it unsuitable for real-time monitoring in a NICU setting.

Stevenson and his colleagues have proposed a similar approach, but based on EEG data, which overcomes this issue. In the original study [46] they used EEG data from 39 preterm infants born before 28 weeks of GA, and who presented normal neurodevelopmental outcome at 12 months of age. The EEG recordings were performed from 24 to 38 weeks PMA. This data was used in a SVM regressor to estimate the infants' PMA. They found that the estimated PMA and the clinically determined PMA were highly correlated (p-value = 0.936), suggesting that the PMA thus estimated could be used as a surrogate measure of the brain maturation. These findings were corroborated in a later study [47] carried out on a bigger population of infants from NICUs of two different hospitals in different countries. The methodology was very similar, with the exception that in this study they also included infants who had an abnormal neurodevelopmental outcome. Once again, they observed a strong correlation between the estimated PMA and the clinically determined one, with a MAE of 0.7 weeks. Furthermore, they reported that a persistently negative difference between estimated PMA and true PMA was associated with poor neurodevelopmental outcome.

Furthermore, in a posterior study [48], Stevenson et al. compared the performance of seven experts on the estimation of the PMA from EEG data, with that of their proposed method, and found that their method provided the most accurate maturity assessment. These three studies establish a proof of concept for growth charts for brain function in NICU, based on EEG data and machine learning, as a new tool to assist clinical evaluation and identification of infants who might benefit from early intervention.

However, it is interesting to observe that there has been a lack of research on the evaluation of maturation of preterm infants based on other physiological signals which are continuously available in the NICU setting, such as heart rate, respiration rate, or apnea and bradycardia events. This is specially noteworthy in the case of heart rate analysis, given that differences in the HRV between healthy preterm neonates and infants born at term have been extensively documented [49], and that these differences persist even when the infants reach term-equivalent age ([50, 51]). Therefore, the exploration of similar techniques to the ones that have been discussed in this section, but relying on other heart rate and respiration rate related data, could prove useful for neonatologists, and help them better assess the health of preterm infants during their

stay in NICU.

1.3 The Digi-NewB Proposal

The Digi-NewB project aimed to answer some of the question left unanswered by current scientific literature in the topic of neonatal care, specifically in regards to LOS diagnosis and objective evaluation of the infants' maturation. This project received founding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 689260, and was carried out between March 2016 and May 2020.

The project proposed the design of a decision support system (DSS) to assist physicians in the early diagnosis and treatment of perinatal infection, and to contribute to limit the risk associated with cardio-respiratory events, intermittent hypoxia, and sleep disturbances. Achieving this could lead to reduce the mortality rate, decrease the risk of neurodevelopmental impairment, shorten the duration of hospitalization, and lower health-care costs.

This DSS would be the result of composite indices that will incorporate and combine available clinical data, and multi-signal analysis, including heart rate, respiration rate, video, and sound signals. These indices would then be able to give adequate, clinically relevant, and continuously updated information, cutting across multiple underlying physiological processes, to aid physicians in the assessment of perinatal health status and development. Also, part of the project's concern is to present these indices in a user friendly interface, that allows interactive interaction with the DSS, so that physicians can easily interpret it, thus increasing usability and acceptability of the system among the healthcare personnel.

The goals of the project are divided in two targets: the main one being the early diagnosis of LOS, and the second one the evaluation of cardio-respiratory, neurobehavioral, and sleep maturation. Specifically, in regards to LOS detection, the Digi-NewB proposal is to identify significant and quantifiable increases in the risk of sepsis, to allow the setting of new, individualized, and accurate evaluation of the evolution of perinatal health, leading to new and earlier therapeutic intervention strategies.

Regarding the maturation objective, the Digi-NewB goal is to develop new, secure, and objective indices that lead to a better informed decision making in terms of when to stop monitoring, to stop assisted ventilation, and to discharge patients from the NICU. The resulting system should also be able to help identify infants who are at a high risk of apparent life threatening events, so that physicians can design specific strategies to monitor this at risk population, specially after stressor events. Finally, it should also help identify infants at significant risk of neurodevelopmental impairment and motor deficits, to allow for early intervention leading to improved long-term outcomes.

One important characteristic of the Digi-NewB proposal, is that it aims to achieve these goals

using only traditionally available data, and sound and video recording. Therefore, it wouldn't require any extra tests or examinations to be done, nor any additional sensors or hardware to be in direct contact with the infants' skin.

In Figure 1.3 we show a real-life example of the Digi-NewB set up for data collection in the NICU of the University Hospital of Rennes. As it can be observed in the picture, the baby is connected only to the usual heart rate monitoring system; cameras and microphones have been set up, to record the movements and sounds of the infants, but these are not in direct contact with the infants. This proposal allows for real-time continuous monitoring, without further disrupting the infants' environment in the NICU, which could negatively impact their neurodevelopmental maturation, and without posing any extra difficulties for the healthcare personnel or for the parents to interact with the infants.

Through these characteristics the Digi-NewB project addresses some of the weaknesses of current practices in the NICUs. For instance, one of the main difficulties currently faced in neonatal diagnosis is that the useful clinical signs and functional analyses are mainly qualitative and are interpreted with a significant amount of subjectivity. Digi-NewB would offer an objective and interpretable measurement derived from these clinical and physiological signs. Another obstacle is that biological markers have a low specificity. However, by applying state-of-the-art machine learning techniques that might help extract the most relevant information from these biomarkers, and facilitate the combination of several signals into one single metric, the Digi-NewB project could offer a higher specificity diagnostic metric. Furthermore, the evaluations of maturity or LOS risk are not continuous and frequently invasive. By using physiological signals that are continuously available, such as ECG, the DSS proposed could be used as a non-invasive and continuously real-time diagnostic tool.

The Digi-NewB proposal could also be an improvement on the already existing AI-based approaches. For instance, by proposing new features for the characterization of the HRV, it might potentially offer better results for early LOS detection than currently accepted and used in monitoring systems, such as HeRO. By integrating different physiological signals into the system, it could offer a more continuous solution than systems such as RALIS. Similarly, on the maturation target, by including ECG, respiration, video, and sound data, it could complement the information given by currently proposed approaches, which are based on the study of the function of the brain through EEGs and MRI.

The scope of the Digi-NewB project was very wide, ranging from data acquisition to the presentation of the resulting monitoring systems in a user-friendly interface, and including many intermediate steps such as data processing and feature extraction, data exploitation through machine learning and data analysis techniques. Furthermore, this process had to be applied simultaneously for several types of data (clinical signs, physiological signals, video image and sound) with very different characteristics.

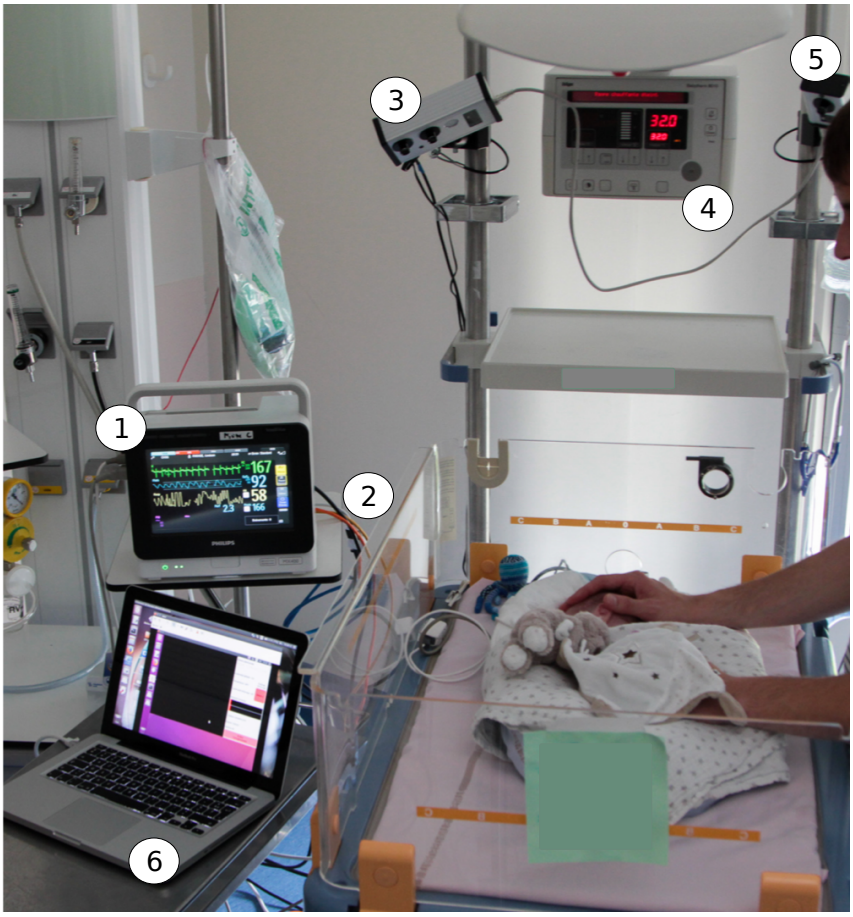


Figure 1.3 – Real-life setting of the Digi-NewB NICU data acquisition system. The elements of interest labeled in this picture are as follows: (1) Vital signs monitor (heart rate, respiration, and oxygen saturation). (2) Digi-NewB cables for data acquisition from the vital signs monitor. (3) Digi-NewB camera and microphone. (4) Radiant warmer control monitor [52]. (5) Digi-NewB camera. (6) Digi-NewB computer for the acquisition and monitoring systems.

Therefore, different teams specialized in certain phases of the data processing or on specific data types. The first stages of data acquisition were handled by physicians and nurses and a team of engineers who handled the data collection from each of the hospitals involved in the study and its transfer to the Signal and Image Processing Laboratory (LTSI, Rennes, France), from which it was then made available to the other partners involved in the project. These data was then processed and exploited by different researches in the project according to the data type and stage of the processing.

In regards to the video data, multiple members of the LTSI were involved in processing the data and exploiting it for several different objectives, including the automated detection of the presence of infants in the incubators [53], and the estimation and characterization of the

neonates' movements [54]. Several studies have been done on the audio data as well, focusing on the automatic detection [55] and tracking [56] of the infants' spontaneous cries. Both video and audio data were also exploited for the analysis of sleep stages in the infants ([57, 58].

Regarding physiological data, algorithms that had been previously developed in the LTSI for the characterization of respiration rate [59] and bradycardia [60] in neonates were used for the extraction of this type of features. For the analysis of the ECG signal [61] and extraction of heart rate variability features [62] new methods were proposed under the frame of the Digi-NewB project. The scope of the work done in this dissertation is a continuation of the work previously done in this regard, and is limited to the exploitation of these features derived from heart rate, respiration rate, and bradycardia signals that were acquired in the Digi-NewB project, through the use of data analysis and machine learning techniques towards the objectives of early LOS detection and maturation evaluation.

1.4 Conclusion

In this chapter we offered a definition of artificial intelligence and machine learning, and explained the basic mechanism of some machine learning algorithms that have been widely used for applications in the field of neonatology, and that will be used in the next chapters of this dissertation. Then, we discussed the state of the art in the field of AI applied to neonatal care, focusing in the current approaches to early LOS diagnosis, and the evaluation of maturation in the preterm infants. Finally, we discussed the Digi-NewB project, in which the work of this thesis is framed, and how it could improve current practices and machine learning approaches in neonatal medicine.

From this discussion, it became apparent that in the area of early LOS diagnosis, improvements could be made by introducing more features that can better characterize changes in heart rate associated with infection. Also, by introducing more complex machine learning models that might be better suited to capture variations in the heart rate over time. Regarding the maturation target, we observed that there is currently a need for models that take into account the physiological signals that are routinely and continuously available in NICU, such as a heart rate and respiration rate. The next chapters of this dissertation will focus on addressing these needs.

Bibliography

[1] J. McCarthy. What is AI? Stanford University. [Online]. Available: http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

[2] S. Russell and P. Norvig, *Artificial intelligence : a modern approach.* Third edition. Upper Saddle River, N.J. : Prentice Hall, ©2010, [2010].

[3] M. Helm, A. Swiergosz, H. Haeberle, J. Karnuta, J. Schaffer, V. Krebs, A. Spitzer, and P. Ramkumar, "Machine learning and artificial intelligence: Definitions, applications, and future directions," *Current Reviews in Musculoskeletal Medicine*, vol. 13, 02 2020.

[4] T. M. Mitchell, *Machine learning.* McGraw Hill Series in Computer Science. Maidenhead: McGraw-Hill, 1997.

[5] K. Kersting, "Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines," *Frontiers in Big Data*, vol. 1, p. 6, 2018.

[6] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database The Journal of Biological Databases and Curation*, vol. 2020, 03 2020.

[7] O. Kramer, *K-Nearest Neighbors.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–23.

[8] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *Journal of Clinical Epidemiology*, vol. 110, pp. 12–22, 2019.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] Y. Qi, *Random Forest for Bioinformatics.* Boston, MA: Springer US, 2012, pp. 307–323.

[11] J. Weston and C. Watkins, "Multi-class support vector machines," Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, Tech. Rep., 1998.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, 2009.

[13] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[14] S. Suthaharan, *Support Vector Machine*. Boston, MA: Springer US, 2016, pp. 207–235.

[15] S. Mirjalili, *Genetic Algorithm*. Cham: Springer International Publishing, 2019, pp. 43–55.

[16] A. K. Jain, Jianchang Mao, and K. M. Mohiuddin, "Artificial neural networks: a tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[17] B. Zamanlooy and M. Mirhassani, "Efficient vlsi implementation of neural networks with hyperbolic tangent activation function," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 1, pp. 39–48, 2014.

[18] P. Petersen and F. Voigtlaender, "Optimal approximation of piecewise smooth functions using deep relu neural networks," *Neural Networks*, vol. 108, pp. 296–330, 2018.

[19] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6.

[20] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015.

[21] A. Abraham, *Artificial Neural Networks*. American Cancer Society, 2005.

[22] Z. Hoodbhoy, S. Masroor Jeelani, A. Aziz, M. I. Habib, B. Iqbal, W. Akmal, K. Siddiqui, B. Hasan, M. Leeflang, and J. K. Das, "Machine learning for child and adolescent health: A systematic review," *Pediatrics*, vol. 147, no. 1, 2021.

[23] I. Daunhawer, S. Kasser, G. Koch, L. Sieber, H. Cakal, J. Tütsch, M. Pfister, S. Wellmann, and J. E. Vogt, "Enhanced early prediction of clinically relevant neonatal hyperbilirubinemia with machine learning," *Pediatric Research*, vol. 86, no. 1, pp. 122–127, 2019.

[24] R. Vassar, K. Schadl, K. Cahill-Rowley, K. Yeom, D. Stevenson, and J. Rose, "Neonatal brain microstructure and machine-learning-based prediction of early language development in children born very preterm," *Pediatric Neurology*, vol. 108, pp. 86–92, 2020.

[25] S. Mani, A. Ozdas, C. Aliferis, H. A. Varol, Q. Chen, R. Carnevale, Y. Chen, J. Romano-Keeler, H. Nian, and J.-H. Weitkamp, "Medical decision support using machine learning for early detection of late-onset neonatal sepsis," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 326–336, 09 2013.

[26] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[27] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[28] M. A. Mazurowski and G. D. Tourassi, "Evaluating classifiers: Relation between area under the receiver operator characteristic curve and overall accuracy," in *2009 International Joint Conference on Neural Networks*, 2009, pp. 2045–2049.

[29] A. J. Masino, M. C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C. P. Bonafide, F. Balamuth, M. Schmatz, and R. W. Grundmeier, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," *PLOS ONE*, vol. 14, no. 2, pp. e0 212 665–, 02 2019.

[30] R. E. Schapire, *Explaining AdaBoost*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 37–52.

[31] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[32] I. Gur, A. Eisenkraft, G. Markel, Y. Nave, D. Bader, and F. Eyal, "Early detection of late-onset sepsis in very low birth weight infants," *Pediatric Research*, vol. 70, no. 5, pp. 72–72, 2011.

[33] I. Gur, A. Riskin, G. Markel, D. Bader, Y. Nave, B. Barzilay, F. G. Eyal, and A. Eisenkraft, "Pilot study of a new mathematical algorithm for early detection of late-onset sepsis in very low-birth-weight infants," *Am J Perinatol*, vol. 32, no. 04, pp. 321–330, 2015.

[34] L. B. Mithal, R. Yogev, H. L. Palac, D. Kaminsky, I. Gur, and K. K. Mestan, "Vital signs analysis algorithm detects inflammatory response in premature infants with late onset sepsis and necrotizing enterocolitis," *Early Human Development*, vol. 117, pp. 83–89, 2018.

[35] K. Fairchild and J. Aschner, "HeRO monitoring to reduce mortality in NICU patients," in *Research and Reports in Neonatology*, 2012.

[36] M. P. Griffin, T. M. O'Shea, E. A. Bissonette, F. E. Harrell, D. E. Lake, and J. R. Moorman, "Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness," *Pediatric Research*, vol. 53, no. 6, pp. 920–926, 2003.

[37] J. R. Moorman, W. A. Carlo, J. Kattwinkel, R. L. Schelonka, P. J. Porcelli, C. T. Navarrete, E. Bancalari, J. L. Aschner, M. Whit Walker, J. A. Perez, C. Palmer, G. J. Stukenborg, D. E. Lake, and T. Michael O'Shea, "Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial," *The Journal of pediatrics*, vol. 159, no. 6, pp. 900–6.e1, 12 2011.

[38] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Frontiers in Public Health*, vol. 5, p. 258, 2017.

[39] S. Nault, V. Creuze, S. Al-Omar, A. Levasseur, C. Nadeau, N. Samson, R. Imane, S. Tremblay, G. Carrault, P. Pladys, and J.-P. Praud, "Cardiorespiratory alterations in a newborn ovine model of systemic inflammation induced by lipopolysaccharide injection," *Frontiers in Physiology*, vol. 11, p. 585, 2020.

[40] S. Saha, A. Pagnozzi, P. Bourgeat, J. M. George, D. Bradford, P. B. Colditz, R. N. Boyd, S. E. Rose, J. Fripp, and K. Pannek, "Predicting motor outcome in preterm infants from very early brain diffusion mri using a deep learning convolutional neural network (cnn) model," *NeuroImage*, vol. 215, p. 116807, 2020.

[41] A. Harel-Gadassi, E. Friedlander, M. Yaari, B. Bar-Oz, S. Eventov-Friedman, D. Mankuta, and N. Yirmiya, "Developmental assessment of preterm infants: Chronological or corrected age?" *Research in Developmental Disabilities*, vol. 80, pp. 35–43, 2018.

[42] K. Schadl, R. Vassar, K. Cahill-Rowley, K. W. Yeom, D. K. Stevenson, and J. Rose, "Prediction of cognitive and motor development in preterm children using exhaustive feature selection and cross-validation of near-term white matter microstructure," *NeuroImage: Clinical*, vol. 17, pp. 667–679, 2018.

[43] K. Cahill-Rowley, K. Schadl, R. Vassar, K. W. Yeom, D. K. Stevenson, and J. Rose, "Prediction of gait impairment in toddlers born preterm from near-term brain microstructure assessed with DTI, using exhaustive feature selection and cross-validation," *Frontiers in Human Neuroscience*, vol. 13, p. 305, 2019.

[44] C. D. Smyser, N. U. F. Dosenbach, T. A. Smyser, A. Z. Snyder, C. E. Rogers, T. E. Inder, B. L. Schlaggar, and J. J. Neil, "Prediction of brain maturity in infants using machine-learning algorithms," *NeuroImage*, vol. 136, pp. 1–9, 08 2016.

[45] P. Galdi, M. Blesa, D. Q. Stoye, G. Sullivan, G. J. Lamb, A. J. Quigley, M. J. Thrippleton, M. E. Bastin, and J. P. Boardman, "Neonatal morphometric similarity mapping for predicting brain age and characterizing neuroanatomic variation associated with preterm birth," *NeuroImage: Clinical*, vol. 25, p. 102195, 2020.

[46] N. J. Stevenson, L. Oberdorfer, N. Koolen, J. M. O'Toole, T. Werther, K. Klebermass-Schrehof, and S. Vanhatalo, "Functional maturation in preterm infants measured by serial recording of cortical activity," *Scientific Reports*, vol. 7, no. 1, p. 12969, 2017.

[47] N. J. Stevenson, L. Oberdorfer, M.-L. Tataranno, M. Breakspear, P. B. Colditz, L. S. de Vries, M. J. N. L. Benders, K. Klebermass-Schrehof, S. Vanhatalo, and J. A. Roberts,

"Automated cot-side tracking of functional brain age in preterm infants," *Annals of Clinical and Translational Neurology*, vol. 7, no. 6, pp. 891–902, 2020.

[48] N. J. Stevenson, M.-L. Tataranno, A. Kaminska, E. Pavlidis, R. R. Clancy, E. Griesmaier, J. A. Roberts, K. Klebermass-Schrehof, and S. Vanhatalo, "Reliability and accuracy of eeg interpretation for estimating age in preterm infants," *Annals of Clinical and Translational Neurology*, vol. 7, no. 9, pp. 1564–1573, 2020.

[49] S. Cardoso, M. J. Silva, and H. Guimarães, "Autonomic nervous system in newborns: a review based on heart rate variability." *Childs Nervous System*, vol. 33, no. 7, pp. 1053–1063, Jul. 2017.

[50] H. Patural, V. Pichot, F. Jaziri, G. Teyssier, J.-M. Gaspoz, F. Roche, and J.-C. Barthelemy, "Autonomic cardiac control of very preterm newborns: a prolonged dysfunction." *Early Human Development*, vol. 84, no. 10, pp. 681–687, Oct. 2008.

[51] E. Helander, N. Khodor, A. Kallonen, A. Värri, H. Patural, G. Carrault, and P. Pladys, "Comparison of linear and non-linear heart rate variability indices between preterm infants at their theoretical term age and full term newborns," in *EMBEC & NBC 2017*, H. Eskola, O. Väisänen, J. Viik, and J. Hyttinen, Eds. Singapore: Springer Singapore, 2018, pp. 153–156.

[52] E. F. Bell, "Infant incubators and radiant warmers." *Early Hum Dev*, vol. 8, no. 3-4, pp. 351–375, Oct 1983.

[53] R. Weber, S. Cabon, A. Simon, F. Poree, and G. Carrault, "Preterm newborn presence detection in incubator and open bed using deep transfer learning." *IEEE J Biomed Health Inform*, vol. PP, Mar 2021.

[54] S. Cabon, F. Porée, A. Simon, M. Ugolin, O. Rosec, G. Carrault, and P. Pladys, "Motion estimation and characterization in premature newborns using long duration video recordings," *IRBM*, vol. 38, no. 4, pp. 207–213, 2017.

[55] S. Cabon, B. Met-Montot, F. Porée, O. Rosec, A. Simon, and G. Carrault, "Automatic extraction of spontaneous cries of preterm newborns in neonatal intensive care units," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1200–1204.

[56] B. Met-Montot, S. Cabon, G. Carrault, and F. Porée, "Spectrogram-based fundamental frequency tracking of spontaneous cries in preterm newborns," in *28th European Signal Processing Conference, EUSIPCO 2020*. European Signal Processing Conference, EUSIPCO, Aug. 2020, pp. 1185–1189. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03129839

[57] S. Cabon, F. Porée, A. Simon, B. Met-Montot, P. Pladys, O. Rosec, N. Nardi, and G. Carrault, "Audio- and video-based estimation of the sleep stages of newborns in neonatal intensive care unit," *Biomedical Signal Processing and Control*, vol. 52, pp. 362–370, 2019.

[58] L. Cailleau, R. Weber, S. Cabon, C. Flamant, J.-M. Roué, G. Favrais, G. Gascoin, A. Thollot, M. Esvan, F. Porée, and P. Pladys, "Quiet sleep organization of very preterm infants is correlated with postnatal maturation," *Frontiers in pediatrics*, vol. 8, pp. 559 658–559 658, 09 2020.

[59] X. Navarro, F. Porée, A. Beuchée, and G. Carrault, "Artifact rejection and cycle detection in immature breathing: Application to the early detection of neonatal sepsis," *Biomedical Signal Processing and Control*, vol. 16, pp. 9 – 16, 2015.

[60] M. Altuve, G. Carrault, A. Beuchée, P. Pladys, and A. I. Hernández, "On-line apnea-bradycardia detection using hidden semi-markov models," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 4374–4377.

[61] M. Doyen, D. Ge, A. Beuchée, G. Carrault, and A. I. Hernández, "Robust, real-time generic detector based on a multi-feature probabilistic method," *PLOS ONE*, vol. 14, no. 10, pp. 1–22, 10 2019.

[62] T. Nguyen Phuc Thu, A. I. Hernández, N. Costet, H. Patural, V. Pichot, G. Carrault, and A. Beuchée, "Improving methodology in heart rate variability analysis for the premature infants: Impact of the time length," *PLOS ONE*, vol. 14, no. 8, pp. 1–14, 08 2019.

CHAPTER 2

Early diagnosis of late onset sepsis in premature infants using visibility graph analysis of heart rate variability

2.1 Introduction

In the previous chapters we discussed the prevalence of late onset sepsis (LOS) in premature infants and the fact that it is one of the main causes of morbidity and mortality in this population [1]. Studies have found that prompt diagnosis and administration of antibiotics can significantly reduce mortality ([2, 3]). However, the indiscriminate use of antibiotics must also be avoided, as it can cause harmful side effects to the patients ([4, 5]) and further increase the level of antimicrobial resistance [6], which is already considered a threat to global public health by the World Health Organization [7].

Therefore, prompt and accurate diagnosis, leading to adequate use of antibiotics is the key to decrease sepsis-related morbidity and mortality, while also protecting patients from unnecessary antibiotic treatment. However, blood cultures and other laboratory tests used to diagnose sepsis are invasive, take time, and present variations in their predictive value, especially in the early phases of infection [8]. Alternatively, changes in the heart rate have been associated with neonatal sepsis and have been suggested as a biomarker for LOS diagnosis. [9].

Heart rate variability (HRV) is defined as the variation of the duration of the interval between consecutive heartbeats over a period of time. Diagnosis relying on heart rate and HRV have the advantage of being non-invasive and readily and continuously available in the context of neonatal intensive care unit (NICU). HRV analysis typically relies on three different categories of features: time-domain, frequency-domain, and non-linear measurements. Previous studies have shown that machine learning algorithms using HRV, based on these features as input, can be useful in early detection of sepsis both in infants and adults ([10, 11, 12, 13]). More recently, network-based time series analysis has also been applied to HRV ([14, 15]), and visibility graph features have shown to be of interest for the diagnosis of different conditions known to alter HRV characteristics ([16, 17]).

Studies have found that some of the features derived from the visibility graph analysis of the HRV time series have a weak correlation to the traditional features, both in adults [18] and infants [19], which suggest that these features might add complementary information to HRV

analysis. One previous study used network analysis of heart rate and blood pressure as input features, alongside multiscale entropy features and clinical measurements from the patients' electronic medical record, for a machine learning algorithm (MLA) that successfully predicted sepsis in adults, achieving an area under the receiver operating characteristics curve (AUROC) of 80% on the test population; an improvement of 7% over the AUROC obtained by their model trained on only the multiscale entropy features and clinical measurements [20].

With the study discussed in this chapter, we aimed to test the diagnostic value of HRV analysis integrating new visibility graph indexes when used in MLAs, in combination with the traditional HRV features, to discriminate between septic and non-septic infants in a selected population of premature infants.

In the following sections, we describe the database used for the study, and the chain of treatment from acquisition of the electrocardiogram (ECG) to computation of the HRV features. We describe the preprocessing of the data and generation of different variants of the feature set, and the MLAs employed to predict sepsis in our population. Finally, we present the evaluation of the machine learning models used on the different variants of the feature set, and then present the results for two sample cases as examples of the differences in HRV between septic and non-septic infants, and the predictions made by the best performing MLA. In the last section we discuss these results and compare them with other results reported in the literature.

2.2 Materials and Methods

2.2.1 Population

The data used in this study is part of the database of the Digi-NewB cohort (NCT02863978, EU GA n°689260). The cohort prospectively included preterm infants born before 30 weeks of gestation, hospitalized in the NICU of six university hospitals in the western region of France (University Hospitals of Rennes, Angers, Nantes, Brest, Poitiers, and Tours) in 2017-2019. The collection of the data was carried out after approval by the ethics committee (CPP Ouest 6-598) and informed parental consent. All the patients with available data having received more than five days of antibiotics, beginning at least 72 hours after birth, were included in the LOS group. The control group consisted of infants who did not receive antibiotics after the first three days of life. For this study, we used data coming from 24 infants who developed LOS, and 25 control infants.

The clinical characteristics of the preterm infants studied is presented in Table 2.1. The results are presented as either median and interquartile range (IQR) or as the number of cases and corresponding percentage. Comparisons between the two populations were performed using Mann-Whitney U test or Chi-squared.

Patients in the LOS group were more premature than in the control group. LOS occurred at

	LOS Group (n=24)	Control Group (n=25)	
Gestational age (weeks)	26.5 (25.3-28)	28 (27-28.5)	p <0.01
Birth weight (g)	840 (740-1025)	1107 (925-1260)	p <0.01
Apgar at 1 minute	8 (5-9)	7 (2-8)	NS
Apgar at 5 minutes	9 (8-10)	9 (8-9)	NS
Male/Female	15/9	12/13	NS
Surfactant	17 (71%)	13 (52%)	NS
Twins	5 (21%)	6 (25%)	NS
Premature rupture of membranes >12h	5 (21%)	7 (28%)	NS
Cesarean section	12 (50%)	15 (60%)	NS
Postnatal age at start of antibiotics (days)	8.4 (5.6-10.5)		
Delay between blood culture and start of antibiotics (hours)	1 (1-2.5)		

Table 2.1 – Characteristics of the study population.

8.4 (5.6-10.5) days after birth, with 17 cases of cocci gram positive bacteria on blood cultures. The LOS group consisted of 17 cases of central line-associated bloodstream infection, five cases of central line-associated infection without positive blood culture, and two clinically suspected infections in patients without central line. The bacteria involved were Staphylococcus Haemolyticus (n=5), Staphylococcus Epidermidis (n=4), Staphylococcus Warnerii (n=3), Staphylococcus Capitis (n=2), Staphylococcus Aureus (n=2) and Enterococcus faecalis (n=1).

2.2.2 Proposed approach

The general approach we propose is described in Figure 2.1. In general terms, we acquired the ECG data from both septic and non-septic patients. This data was processed to detect the R peak and thus obtain the R-R peak interval time series, which was then segmented in periods of 30 minutes. The HRV features were then extracted from each of these periods.

Afterwards there was a labeling process for each neonate, in which the first hours of measurements were used as a calibration period, to which the measurements from the remaining hours are compared. Then the hours before the beginning of antibiotic therapy, in the case of septic infants, were labeled as infected. For the control population, we randomly selected a time between the third and tenth day of life (corresponding to the time of late onset sepsis diagnosis in the LOS group), and the hours before it were labeled as no-sepsis.

Different methods were then used to select the features to be passed as input to four different MLAs. We used the leave-one-out cross-validation (LOOC) method, leaving one infant out in each iteration, so each infant was at some point the test patient, while the rest formed the training set. We used 8-fold cross validation, grouped by patients, in the training set to optimize

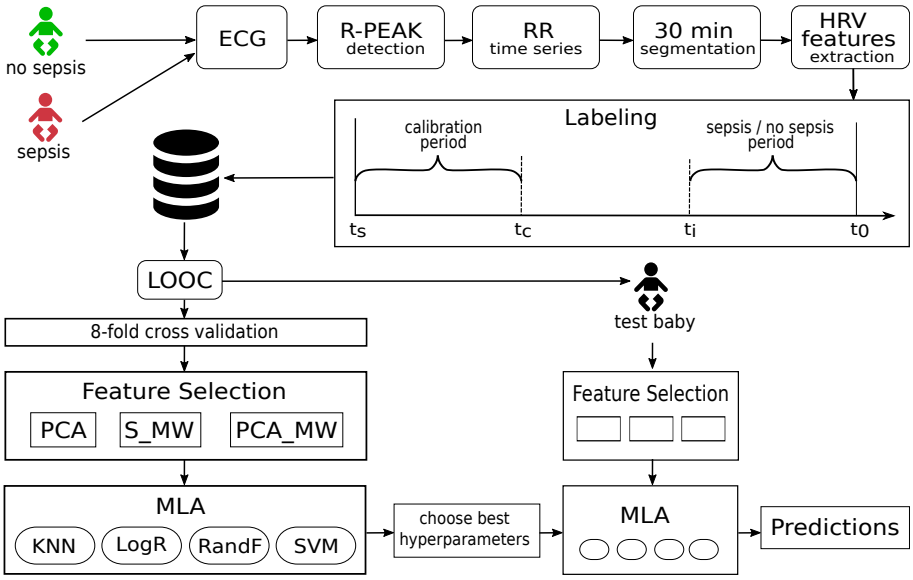


Figure 2.1 – Proposed approach.

the hyperparemeters for each MLA.

The following sections will explain in greater detail each of the steps of our proposed approach.

2.2.3 Signal Processing

The ECG from the infants were obtained with a sampling rate of 500Hz. The RR intervals were detected using a modified version of the Pan and Tompkins algorithm, with filter coefficients specifically adapted for newborns, as proposed in [21]. Afterwards, a sliding window of 30 minutes, with no overlap, was applied to calculate the RR series and from it all the HRV parameters that will be described in section 2.2.4.

2.2.4 Extraction and Analysis of HRV Parameters

Time-Domain Measurements The time domain parameters calculated for this study were the mean of the RR intervals (meanRR), the standard deviation (sdRR), the root mean square of successive RR intervals (RMSSD), the maximum and minimum value for the RR intervals in the time series (maxRR and minRR, respectively)[22], skewness, kurtosis [23], and AC and DC, which characterize the acceleration and deceleration capacity, respectively, of the heart rate [24].

Frequency-Domain Measurements In the frequency domain we calculated the low frequency power (LF), with limits 0.02-0.2Hz, the high frequency power (HF), with limits 0.2-2Hz.

We also calculated these features in normalized units (LFnu and HFnu, respectively), and the LF/HF ratio [22].

Non-linear Measurements The non-linear parameters include the sample entropy (SampEn) and approximate entropy (ApEn), which estimates the level of regularity and predictability of the signal; the coefficients $\alpha 1$ and $\alpha 2$, obtained from the detrended fluctuation analysis of the time series, and which represent, respectively, the short-range and long-range fractal correlations of the signal; and the parameters SD1 and SD2, derived from the Poincaré plot, and which reflect the short and long term variability, respectively [22].

Visibility Graph Indexes The visibility graph (VG) is a network-based time series analysis; it converts the series into a graph that inherits several of its properties, by transforming every data point of the time series into a node. For this we used the visibility graph criterion proposed by Lacasa et al. in [25]. The horizontal visibility graph (HVG) is a subset of the VG, which we calculated using the HVG criterion proposed by Luque et al. in [26].

Several indexes were computed to characterize each graph:

— The mean degree (MD_VG and MD_HVG, respectively) of all the nodes in the graph, where the degree of a node is defined as the number of connected edges of the node. The mean degree is a measure of the complexity of the network [27].

— The cluster coefficient (C_VG and C_HVG, respectively) is the average of the local cluster coefficient of all the nodes in the graph, where the coefficient of a node is a measure of how much its neighbors are also connected to each other, and is defined as the ratio between all triangles involving that node, and the number of connected triples centered on that node [27].

— The transitivity (Tr_VG and Tr_HVG, respectively) is a global version of the cluster coefficient, and is obtained as the ratio between the number of triangles in the graph, and the number of connected triples [27].

— The assortativity (r_VG and r_HVG), which is a correlation between the degrees of all nodes on two opposite ends of an edge, with a graph being assortative if the connected nodes have comparable degree ($r > 0$), and disassortative otherwise ($r < 0$) [28].

The details of how the graphs were constructed and how the indexes were calculated are presented in appendix 2.A.

2.2.5 Data Analysis and Machine Learning

To prepare the data for analysis and machine learning, we first excluded all the 30-minute segments with a maxRR greater than one second, or a minRR of less than 0.19 seconds. Afterwards, for the infected infants we selected all the remaining segments prior to the time of

administration of antibiotics, which is the time we use as the time of LOS onset (t_0); for the infants in the control group we selected a moment at random between the third and tenth day of recording as t_0 (due to the fact that for our LOS group the median value for infection onset is eight days after birth), and selected all the segments prior to that moment. Then, for each infant we calculated the median value of each parameter over a calibration period; we tried three different lengths of calibration period: (i) the first 24 hours of recording, (ii) the first 48 hours of recording, and (iii) the first 72 hours of recording. For the rest of the 30-minute segments of each infant, we subtracted from each feature the median value of that feature obtained over the calibration period. In this way we obtained the calibrated features (Δfeatures). Thus, we generated three variants of the feature set, one for each different length of the calibration period. Additional variants of the feature set were obtained by changing the time we considered as the learning window (t_i): we considered all cases starting from t_i = -72 hours (that is, 72 hours before t_0) [29], until t_i = -6 hours, with six hours increments. For the infected group we labeled the entire duration of the learning window as infected, and for the control group as not infected. Finally, for each variant of the feature set we considered two cases: one including the visibility graph features, and one were they were excluded.

For the statistical analysis and machine learning process we used the LOOC method, leaving one patient out in each iteration. For each variant of the feature set, we used the Mann-Whitney U test to compare the LOS and control population in the training set on each HRV feature, and thus retain only the features that yielded a p-value under 0.1 (MW). Then, we performed principal component analysis (PCA) on the training data with all the features, as well as PCA with only the features with p-value < 0.1 (PCA_MW); in both cases we retained the components for 95% of the variance of the feature set. Thus, we created two different sets of features based on PCA. We created a third set with only the features for which p-value < 0.1, which were standardized (S_MW) before being passed to the MLAs.

We used each of these sets of features to train four different machine learning algorithms: k-nearest neighbors (KNN), logistic regression (LogR), random forest (RandF), and support vector machine (SVM). We used grid search with a per subject 8-fold cross validation split in the training set to find the best hyperparameters specific to each algorithm to maximize recall for each MLA. The list of the hyperparameters tested, as well as the best hyperparameters for each algorithm, are presented in the appendix 2.B.

Finally, each algorithm trained with the best hyperparameters was tested on the patient left out. This process was repeated until every infant in the database had been used as the test subject (the patient left out). The probability curve thus obtained for every patient was then smoothed by calculating for each point in time (equivalent to a 30-minute segment), the probability of infection as the median value between the current predicted probability, and the probability of the two previous segments.

2.2.6 Evaluation Method

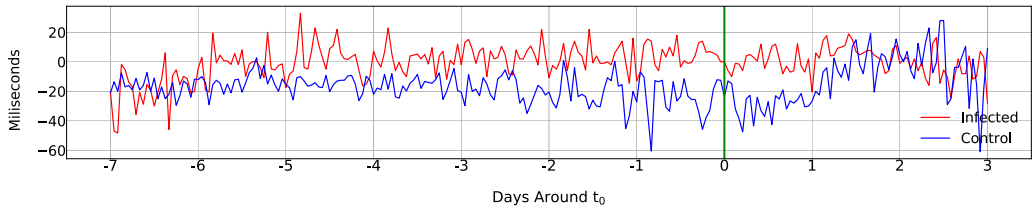
The evaluation of the performance of each MLA has been done in terms of its area under the receiver operating characteristics curve (AUROC). Our main analysis focused on the performance on the time window comprising the six hours before t_0. However, we were also interested in evaluating the performance during earlier periods of time to determine how early the infection could be detected by the MLAs tested. For this purpose, we evaluated the AUROC on a sliding window with a duration of six hours, starting at the interval between -6h to t_0, and ending at -48h to -42h, sliding with a 50% overlap. For both analyses the AUROC was calculated when combining the predictions made for each patient. Confidence intervals for the AUROCs were calculated using the methods and R library described in [30].

To analyse the value added by the visibility graph indexes to the models, we performed a likelihood ratio test ([31, 32]) on the best performing model, to compare its performance when the visibility graph indexes are included in the feature set to when they are not included. With this test, a p-value under 0.05 means that the information added to the model by the new features causes a statistically significant improvement on the model.

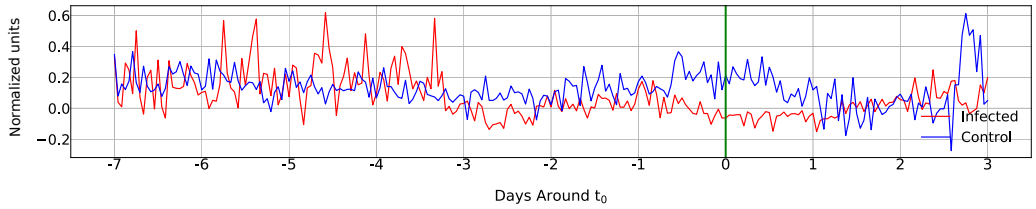
For the purpose of analysis and visualization of the predictions given by a particular MLA, trained on a given variant of the feature set, and for a specific patient, we compared the predicted probability of infection to a fixed threshold of 0.5 for all MLAs, although this might not be the optimal threshold for that MLA. We chose this method to simplify the comparison between the results from different MLAs, by comparing them all based on a set threshold. Thus we considered a false negative as a probability lower than 0.5 for an infected infant, and a false positive as a probability greater than 0.5 for a control patient.

2.3 Results

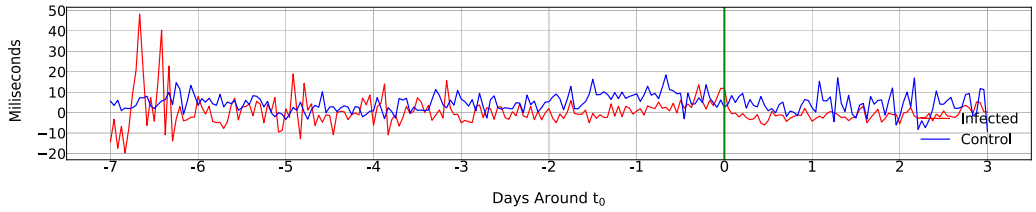
In this section we first present the behavior of certain measurements of the HRV analysis in the whole population. Afterwards we report the predictive performance of the MLAs with the different variants of the feature set. For simplicity, we only present the best results obtained. Next, we present the HRV measurements which showed statistically significant differences between LOS and control group, and then we analyze the effect of varying the calibration and learning windows on the predictive performance of the algorithms. Subsequently we consider the effect of including the visibility graph indexes. Finally we present the results for two patients of our population, one from the control group and one from the LOS group, as sample cases and to exemplify how our method could be used for monitoring in the NICU.



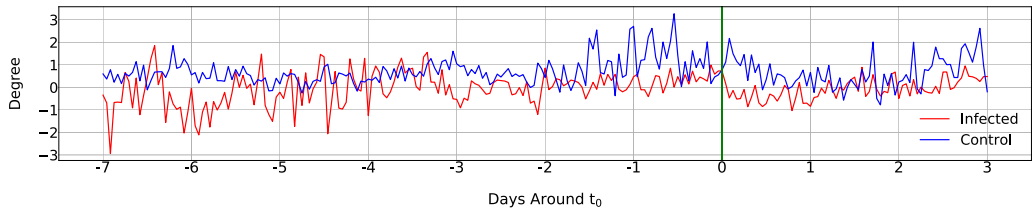
(a) ΔminRR.



(b) ΔLFnu.



(c) ΔSD2.



(d) ΔMD_VG.

Figure 2.2 – Median value of the Δfeatures over several days.

2.3.1 General Behaviour of Some HRV Parameters

In Figure 2.2 we present the comparison between the median values of the 24 infected infants (in red) and the 25 control infants (in blue) for some of the calibrated HRV parameters (Δfeatures), for which a difference between both groups was easily observable. The green line represents the time t_0. The features shown in the figure were calculated using a calibration period of 48 hours.

Figure 2.2a shows the median value of the ΔminRR for the control and sepsis groups, and it can be observed that the value is generally higher for the LOS group, which is consistent with an expected increase of the occurrence of bradycardia in the infected patients. In Figure 2.2b we observe that the ΔLFnu is generally lower in the infected population in the days around the

t_0, as compared to the control group; this could reflect alterations in baroreflex induced changes in HRV during sepsis. The ΔSD2 is also generally lower in the infected group, as observed in Figure 2.2c, which correlates to the decreased HRV associated with sepsis. Similarly, Figure 2.2d also shows evidence of a decrease of the ΔMD_VG in the infected infants, which also signals decreased HRV in this group.

2.3.2 Predictive Performance of the MLAs

We evaluated the predictive performance of all the MLAs using every variant of the feature set, created as explained in section 2.2.5, for a time window comprising the six hours before t_0. These results are presented in Table 2.2, where we show the variant of the feature set that gave the best result, in terms of greatest AUROC, for each of the four types of MLAs we tested. The upper part of the table shows the results when the visibility graph indexes were included in the feature set, while the lower half displays the results obtained when these features were excluded. The *Features* column specifies which feature selection technique was used to construct the feature set (PCA, S_MW, or PCA_MW); the *Calibration* column indicates how many hours were used for the *calibration period*; the t_i column shows how many hours before t_0 were used for the training window. Finally, in the *AUROC* column we present the AUROC, and its corresponding 95% confidence interval (CI), achieved by the algorithm using the given variant of the feature set on both the training and testing data.

In Table 2.2 we observe that the best performance for detecting whether a test patient is infected or not during the six hours before t_0 was obtained by the LogR algorithm, on the feature set that included the visibility graph indexes, with 87.7% AUROC on the testing data.

MLA	**Including Visibility Graph Features**				
	Feature Set Variation			AUROC (% [95% CI])	
	Feature Selection	Calibration	t_i	Training	Testing
KNN	S_MW	48h	-30h	89 [88.7, 89.3]	77.7 [73.1, 82.3]
LogR	PCA_MW	48h	-42h	88.4 [87.9, 88.9]	87.7 [83.3, 92.2]
RandF	PCA_MW	48h	-6h	99 [98.6, 99.4]	81 [75.7, 86.3]
SVM	PCA	72h	-12h	91.5 [90.4, 92.6]	78.3 [72.1, 84.5]
MLA	**Excluding Visibility Graph Features**				
	Feature Set Variation			AUROC (% [95% CI])	
	Feature Selection	Calibration	t_i	Training	Testing
KNN	PCA	48h	-24h	87.9 [87.5, 88.3]	73.2 [66.9, 79.5]
LogR	PCA_MW	48h	-42h	82.2 [81.7, 82.7]	80.9 [75.9, 85.9]
RandF	PCA_MW	48h	-6h	98.4 [97.8, 99]	73 [66.4, 79.6]
SVM	PCA	72h	-6h	92.6 [91.7, 93.3]	82.3 [77.4, 87.2]

Table 2.2 – Best feature set variants for the six hours evaluation window and their respective AUROC. AUROC are presented as median value and 95% confidence interval.

All the MLAs performed better when the visibility graph indexes were included in the feature set, except the SVM which performed better without the visibility graph indexes. Similarly, all MLAs performed better when the calibration period used was of 48 hours, except the SVM which performed better with a calibration period of 72 hours. This might be due to a more robust calibration when using a 72h period, making it less sensitive to the HRV changes normally associated with the first hours of life in neonates [33]; SVM might benefit from this more than the other MLAs given its sensitivity to any noise or outliers in the training data [34].

Regarding the feature selection, LogR and RandF performed better when trained with the PCA_MW features, while KNN performed better on the set of S_MW features, and SVM had a better performance on the PCA of all the features. Finally, the best training window for KNN was 30 hours before t_0, for LogR it was 42 hours before t_0, while for RandF and SVM the best training window was when t_i = -6 hours.

We also observe that, as expected, the AUROCs obtained from the predictions for the training data are bigger than those obtained from the predictions for the testing data. For the KNN, RandF, and SVM the difference between training and testing AUROC ranges from 11.3% to 25.4%. Instead, for the LogR the difference between the training and testing AUROCs is smaller, at 0.7% when visibility graph features are included, and 1.3% when they are not. This suggests that the main reason why the LogR, despite being one of the simplest of the MLAs we tested, outperforms more powerful algorithms such as RandF and SVM, is because it is not over fitting on the training data, while the other algorithms are.

In Figure 2.3 we show how the AUROC from the testing data changes for each algorithm when evaluated on a sliding window of six hours, with a 3-hour overlap, between t_0 and t_0 = -48 hours. For this, each algorithm was evaluated using its optimal variant of the feature set, as presented in Table 2.2. We observe that all algorithms have an AUROC above 60% for all time windows. Furthermore, LogR, RandF, and SVM have an AUROC above 70% since at least 42 hours before t_0. In general, their AUROCs begin rising at 24 hours before t_0 and until t_0, with LogR, RandF, and SVM ending with AUROCs of over 80%. We observe that the AUROC for all algorithms present some oscillations over time, with the LogR being the most stable of the four. This suggests that the oscillations might be due to the overfitting on the training data which, as observed in Table 2.2, was more marked in the other three MLAs, and overfitting can cause small changes in the test data to translate into significant changes in the predictions made.

When visibility graph indexes were excluded, the LogR model had a very similar performance to the SVM, which was the best performing MLA in this case, with a difference in AUROC of only 1.4%, and similar confidence intervals. Also, LogR is a simpler algorithm, faster to train than the others, less prone to overfitting, and it outperformed all other algorithms when the visibility graph features were included. for these reasons, from this point on we focus our results and analysis concerning feature selection and the effect of visibility graph indexes on the results

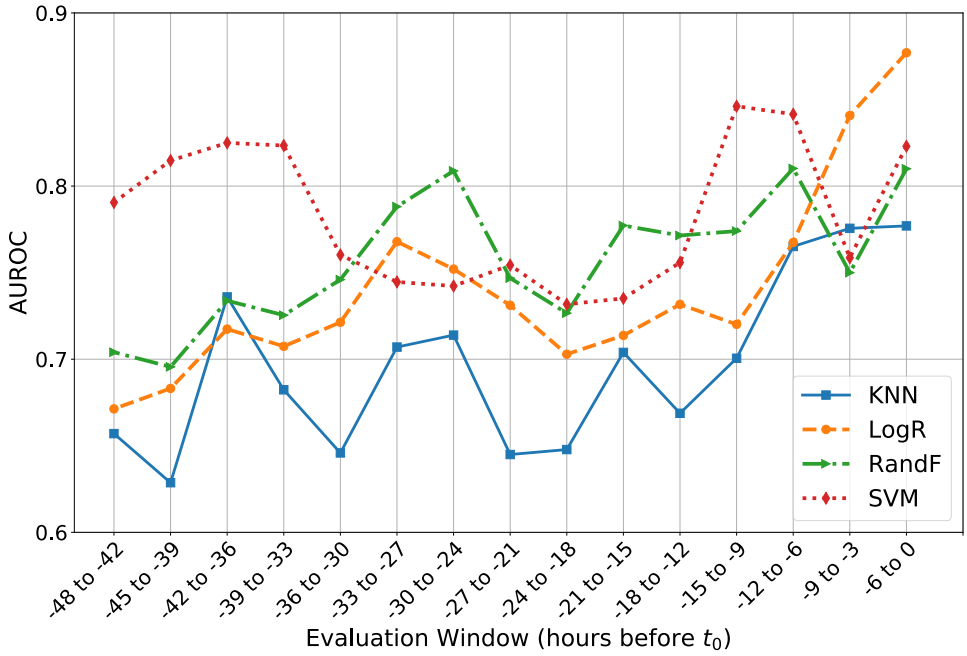


Figure 2.3 – Progress of the AUROC for the best performing MLAs configurations evaluated on a sliding time window.

obtained with the LogR algorithm, using the feature set variant as presented in Table 2.2.

2.3.3 Feature Selection

For every variant of the feature set, we also implemented the Mann-Whitney U test to compare the HRV calibrated features (Δfeatures) of the control group to those of the infected group. We applied this to the training data in every iteration of the LOOC procedure. For simplicity, in this section we present the measurements that had p-value < 0.1, when the calibration period was set at 48h, and the learning window at -42h, as presented in Table 2.2 for the best performing MLA, LogR. In Table 2.3 we present the HRV features which had statistically significant differences (p-value < 0.1) between LOS and control group in at least 50% of the LOOC iterations for said configuration. The column *Occurrence* shows the percentage of the LOOC iterations for which the given measurement showed statistically significant differences.

We observe that out of 28 HRV measurements considered, 24 were relevant (p-value < 0.1) in at least 50% of the cases, when using a calibration window of 48h and a learning window of -42h. Particularly, from the visibility graph features, four were relevant in 100% of the iterations: *MD_VG*, *Tr_VG*, *r_VG*, and *MD_HVG*.

HRV Feature	Occurrence	HRV Feature	Occurrence
meanRR	100%	HFnu	100%
sdRR	96.4%	LF/HF	100%
RMSSD	100%	SD1	100%
maxRR	98.2%	SD2	96.4%
minRR	100%	SampEn	100%
Skewness	100%	ApEn	94.6%
Kurtosis	100%	$\alpha 1$	73.2%
AC	98.2%	$\alpha 2$	100%
DC	98.2%	MD_VG	100%
LF	100%	Tr_VG	100%
HF	98.2%	r_VG	100%
LFnu	100%	MD_HVG	100%

Table 2.3 – HRV measurements with statistically significant differences (p-value < 0.1) between control and infected population. The column Occurrence provides the percentage of LOOC iterations for which the feature is significant.

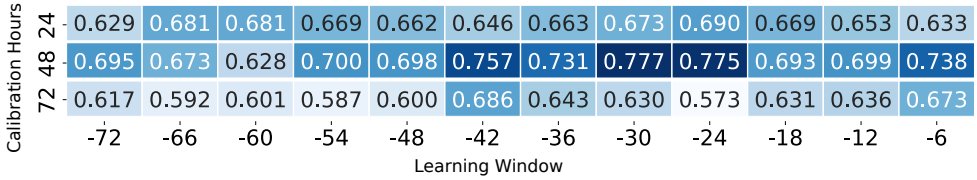
2.3.4 Optimization of the Calibration Period and Learning Window

To analyze the effect of varying the calibration window and learning hours, we evaluated the AUROC for each possible combination when evaluated in the period of six hours before t_0. We did this for each MLA separately, and training and testing them on datasets built using the feature selection method that gave the best result for that MLA, which were shown in Table 2.2. The results for this analysis are presented in Figure 2.4.

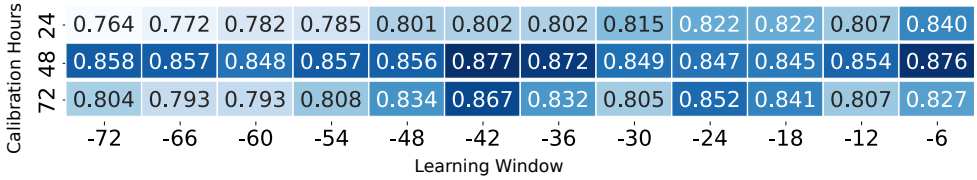
For the KNN, using S_MW as the feature selection, the results are shown in Figure 2.4a. We observe that the best results are obtained when using 48 hours for calibration, and t_i between -42 and -24 hours. Although setting t_i to -54 and -6 hours, still using 48 hours for calibration, also gave results of at least 70%. For all the other configurations the AUROC remained below 70%, but greater than 57%.

In Figure 2.4b we present the results for the LogR algorithm, using the PCA_MW method for feature selection for the training and test datasets. In this case the best results were also mostly obtained when using the 48 hours calibration period, with the minimum AUROC for that case being of 84.8%. Using 72 hours of calibration also gave good results; the best AUROC for this case was obtained when using -42 hours in the learning window, at 86.7%, only 1% below the best AUROC for this algorithm, which was also obtained with t_i = -42h, but calibration hours set to 48. Using the calibration period of 24 hours resulted in the lowest AUROCs for the LogR. The lowest AUROC for this MLA was of 76.4%, obtained when using 24 hours for calibration and -72 hours for the learning window.

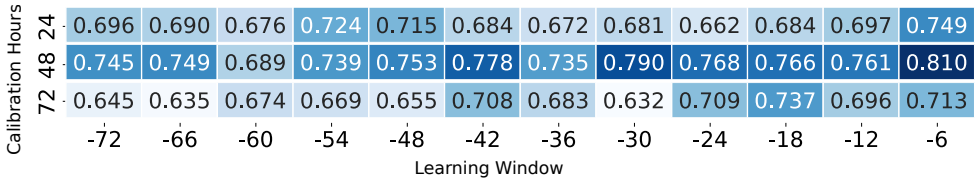
The performance of the RandF algorithm was analyzed using the PCA_MW feature selection method for the training and testing datasets, and the results are presented in Figure 2.4c. In this



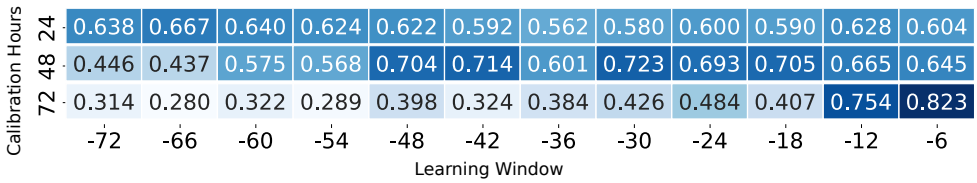
(a) K-Nearest Neighbors.



(b) Logistic Regression.



(c) Random Forest.



(d) Support Vector Machine.

Figure 2.4 – Variations of the AUROC of the MLAs as the calibration hours and learning windows change.

case the best results were also obtained using 48 hours before calibration and, in general, with the learning window between -42 and -6, with the best case of all being 48 hours for calibration and -6 hours for learning, which yielded an AUROC of 81%. In general, this algorithm performed worse when using 72 hours for calibration, with the lowest AUROC (63.2%) obtained when using t_i = -30 hours, and 72 hours as the calibration period.

For the SVM we used the PCA feature selection method for analyzing the performance of the algorithm, as the calibration and learning windows varied. These results are shown in Figure 2.4d. For this algorithm we observed the greatest variation in the AUROC; even though SVM gave the second best overall result of all the algorithms (82.3%, when using 72 hours for calibration and -6 hours in the learning window), it is also the only algorithm that had AUROCs under 50%, with the worst performance being obtained using 72 hours for calibration and -66 hours for learning, which yielded an AUROC of 28%. For this MLA, in fact, we observed an inverse

relation between calibration hours and learning window: with a 24 hour calibration period, it performed better with a wide learning window of between -72 to -60 hours; when the calibration period increased to 48 hours, it performed better in the intermediate values of the learning window, getting AUROCs over 70% with t_i set to -48, -42, -30, and -18 hours; instead, with the calibration window set to 72 hours, all learning windows wider than 12 hours had AUROCs under 50%, but the best results for this algorithm were obtained using the calibration period of 72 and learning windows of -12 and -6 hours (AUROCS of 74.5% and 82.3%, respectively).

2.3.5 Effect of Visibility Graph Indexes

As it was shown in Table 2.2, the MLA with the best predictive performance for the six hours before the infection was the LogR, using a calibration period of 48 hours and the time for the onset of the infection set at 42 hours before the administration of antibiotics, using as input features the principal components of the features with p-value < 0.1 (PCA_MW), and including the visibility graph indexes in the feature set. This setting obtained an AUROC of 87.7%, which is presented as the blue solid line (Visibility) in Figure 2.5.

When excluding the visibility graph features, the LogR model, using the same calibration period, learning window, and feature selection method as before, obtained an AUROC of 80.9%, which is presented as the green dotted line (No Visibility) in Figure 2.5.

Thus introducing the visibility graph indexes in the feature set increased the performance of the MLA by 6.8%. Furthermore, we performed a likelihood ratio test to determine if the improvement obtained by the inclusion of the the visibility graph indexes to the feature set is statistically significant, obtaining a p-value of 3.5e−6 in the training set, and 2.9e−4 in the testing set. This indicates that the information added by the visibility graph leads to an statistically significant improvement in the fit of the model.

2.3.6 Sample Cases

The two cases presented in this section illustrate the results obtained for two patients: one from the LOS group and one from the control group. The patient from the control group was an extreme preterm male, born at 27 weeks gestation, with birth weight of 1220g and Apgar score of 2/8, and received 48 hours of antibiotics at birth for an unconfirmed suspicion of early onset sepsis. He was treated with caffeine and nasal continuous positive ventilation with 23% oxygen. He did not develop any infection during his stay in neonatology.

The patient from the LOS group developed LOS with positive blood culture which identified an Enterococcus Faecalis on the 14th day of life. This patient was also an extreme preterm male, born at 25 weeks of gestation, with birth weight of 730g and Apgar score of 5/7. The patient received 48 hours of antibiotics at birth for an unconfirmed suspicion of early onset sepsis. He was treated with caffeine and nasal intermittent positive ventilation with 28% of oxygen, and

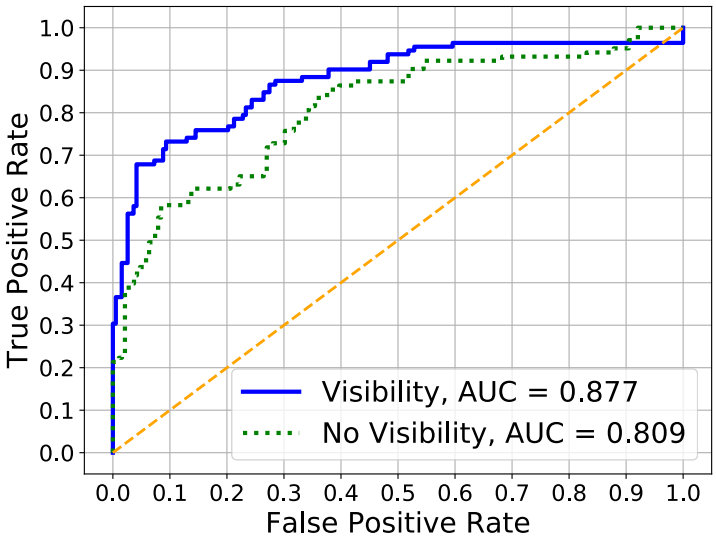
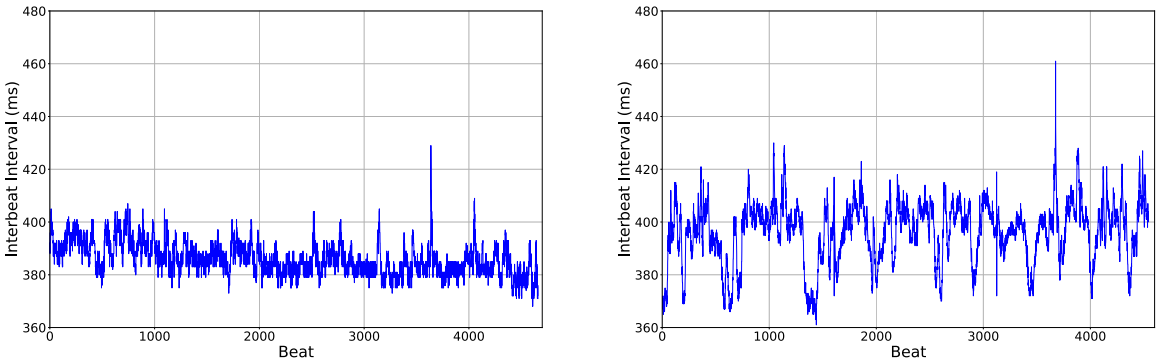


Figure 2.5 – Best predictive performance with and without visibility graph indexes six hours before administration of antibiotics.



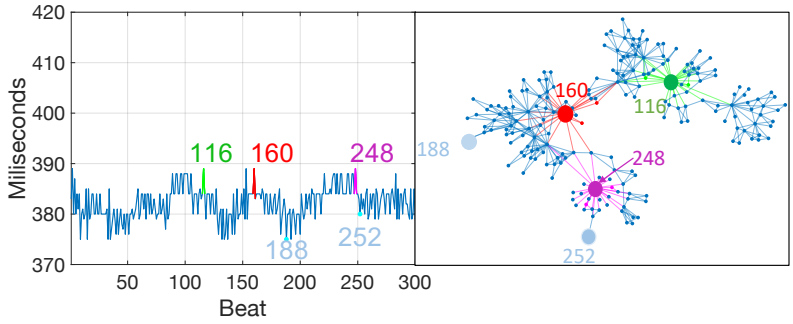
(a) Sepsis patient.

(b) Non-sepsis patient.

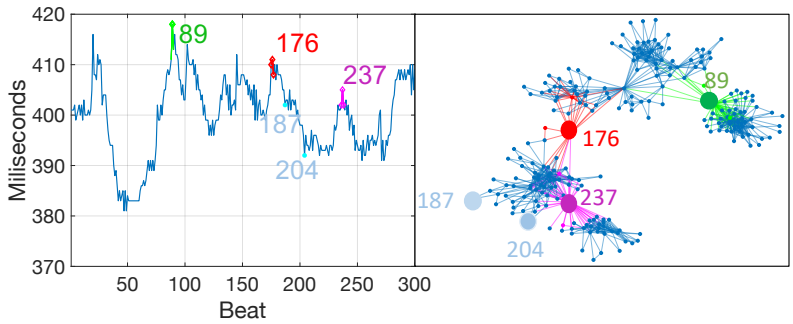
Figure 2.6 – RR time series for 30-minute segments observed three hours before t_0.

fed through a venous central line. At the time of clinical suspicion of LOS an isolated increase in cardio-respiratory events was observed without other clinical signs. The results below show that the proposed method would have been able to diagnose the emerging infection at least 12 hours before the clinical suspicion.

In Figure 2.6a we observe the RR time series from the LOS patient, corresponding to a period of 30 minutes, three hours before the beginning of administration of antibiotics. In Figure 2.6b we present the RR time series of a 30-minute segment from the control patient. In this figure a difference we observe that the infected patient displays less variability in its heart rate in comparison with the patient from the control group.



(a) Sepsis patient.

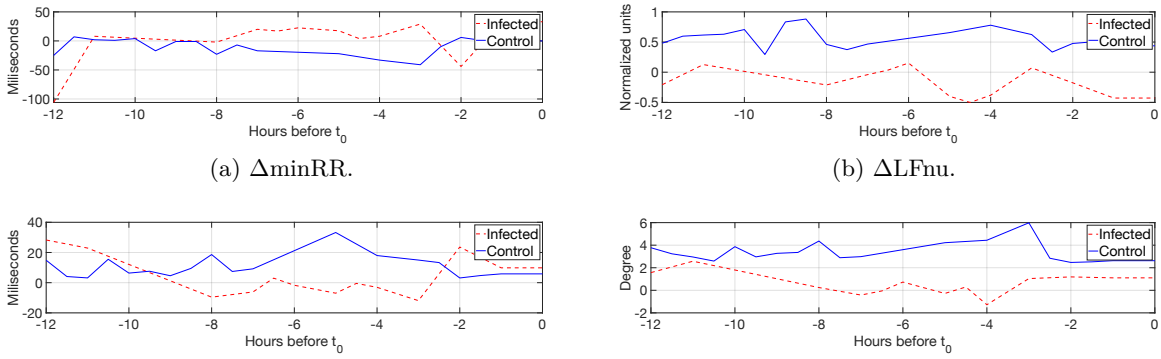


(b) Non-sepsis patient.

Figure 2.7 – RR time series and its corresponding visibility graph

In Figure 2.7 we present the visibility graph obtained from the RR time series presented in Figure 2.6. To facilitate the visual interpretation of the graph, for both patients we have zoomed into a window of only 300 beats of the time series, which is shown in the left side of the figures; we have also highlighted some interesting beats and their respective nodes in the visibility graph: in green, red, and magenta we highlight local maxima of the RR time series, and in grey local minima. We observe that the beats that are local minima in the time series, in the visibility graph convert into nodes that have very few connections and that are in the outer part of the clusters. On the other hand, the nodes that are local maxima convert into nodes that connect different clusters. In Figure 2.7a we observe that the low heart rate variability of the infected baby translates into a visibility graph with less connections within clusters. In comparison, in Figure 2.7b we observe that for the non-septic patient, the connections within each cluster are denser.

The HRV measurements of these patients were calculated and calibrated. Thus, we obtained the features that would be used by the MLA, some of which are shown in Figure 2.8, where we compare the Δfeatures for the same non-septic (blue solid line) and septic (red dashed line) patients shown in Figure 2.6, during the 12 hours before (t_0). Differences in the HRV of the non-septic and LOS patient can be observed in the four different types of measurements



(a) ΔminRR. (b) ΔLFnu.

(c) ΔSD2. (d) ΔMD_VG.

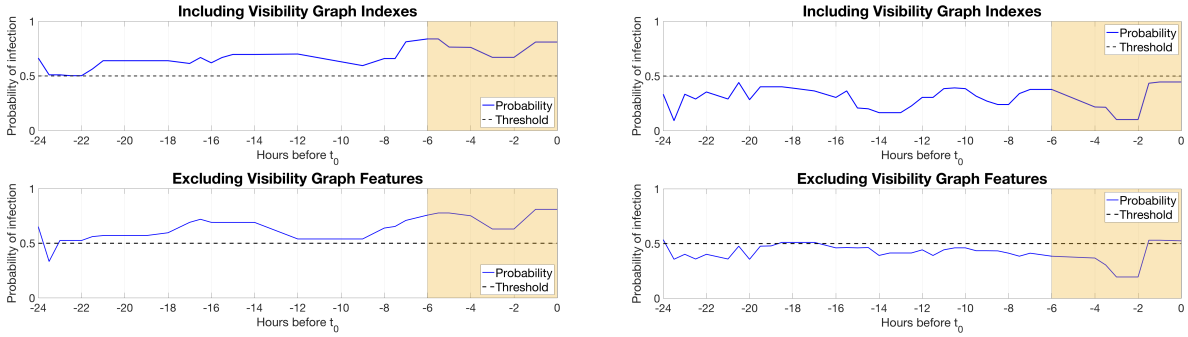
Figure 2.8 – Calibrated Features for Sample Cases.

we considered: time-domain (exemplified by minRR in Figure 2.8a), frequency-domain (LFnu in Figure 2.8b), non-linear measurements (SD2 in Figure 2.8c), and visibility graph indexes (MD_VG in Figure 2.8d). It's important to note that the Δfeatures shown in Figure 2.8 are the same as those shown in Figure 2.2, and that the features for these two sample patients follow the same tendency observed when comparing the entire LOS group to the control group.

Finally, the predictions of the probability of infection over time for both sample patients were calculated using the LogR model, with the configuration which yielded the best results in the six hours before t_0, and including and excluding the visibility graph indexes, as presented in section 2.3.2. In Figure 2.9 we present the predicted probability during the 24 hours before t_0, and highlight in yellow the period corresponding to the six hours before t_0.

The results for the infected infant are presented in Figure 2.9a, with the top row showing the predicted probability (in blue) when the visibility graph features were included, and the bottom row presenting the predicted probability (in blue) when these features were excluded from the feature set. The black dotted line represents the threshold probability of 0.5. In the case where the visibility graph indexes are included, observe that while the probability gets very close to the threshold in the period between -24h and -22h, it never actually crosses the line, so there are no false negatives. On the other hand, when these features are excluded there is a false negative in the period between -24h and -22h.

In the case of the patient from the control group, presented in Figure 2.9b, when visibility graph indexes were excluded there were false positives in the predicted probability of infection (bottom figure), which was not the case when these features were included in the feature set (top figure).



(a) Septic patient.

(b) Non-septic patient.

Figure 2.9 – Predictions six hours before t_0.

2.4 Discussion

This study proposes a method for estimating the probability of LOS in premature neonates using MLAs, with the aim to aid an earlier and more accurate diagnosis. Our proposed method is based on extracting HRV features from the continuous heart rate (HR) monitoring, and then using those features as input for the MLA. For this we use the traditional HRV features: time-domain, frequency-domain, and non-linear measurements. However, we also propose the inclusion of more novel measurements based on visibility graph indexes.

Previous studies have demonstrated that MLAs can detect sepsis in both adults and infants. Unlike the method we have proposed here, most of these studies rely not only on HR measurements, but may also include respiratory rate, blood pressure, motion, clinical signs, and laboratory tests ([35, 36, 37, 38, 29, 39, 40, 41]). And among those studies that have focused exclusively on HR or HRV measurements ([12, 10, 11]), we did not find any study that included visibility graph indexes or network-based analysis. In fact, we could find only one previous study that used visibility graph indexes of HR and blood pressure to diagnose sepsis in adults using MLA, which found a improvement of 7% in the AUROC when these measurements were included [20]. But to our knowledge, no previous study has used visibility graph indexes for diagnosis of sepsis on premature infants.

Thus, one of our main findings was the contribution of visibility graph features to the performance of the MLAs. We found that the AUROC of the best performing MLA improved by 6.8% when the visibility graph indexes were included. The likelihood ratio suggests that the improvement introduced by these features is statistically significant, with p-value $< 3e{-}4$.

Another important aspect of our method is that we use the median value of HRV features for each individual patient during a calibration period as a baseline reference for that patient, and it is the difference between this reference value and the value measured for the following segments what is actually passed as input to the MLA. A similar approach has been proposed

before, both in adults, where one study reports using the mean value of the HRV metrics over the first 24 hours of recording as reference value [42], and in premature infants, where another study used a calibration period of 72 hours to predict sepsis based on HR, respiratory rate, and clinical signs [29], however this study does not specify how the calibration was performed. But this type of method might be specially useful in the case of evaluating HRV in premature infants, for different studies have shown that differences in gestational age can imply significant differences in the HRV of the infant. In our study the best results were obtained when using the first 48 hours of recording as the calibration period.

In regards to the different MLAs evaluated, we found that LogR had a better performance than the others, even though it was also the simplest one used. This might be due to the fact that we did not have a very large population, and precisely due to its simplicity, LogR is less likely to overfit on the training set. This is supported by that fact that all the other MLAs had a significantly bigger AUROC on the training data than on the testing data, while this difference was very small with the LogR. However, with a larger population better results could possibly be obtained using more complex MLAs.

The fact that best performance was achieved when training the MLA using the 42 hours before t_0, for both infected and control patients, might be due to this yielding a bigger dataset for the training of the MLA.

Concerning the preprocessing for feature selection, we obtained the best results when choosing the measurements for which the comparison between LOS and control population had p-value under 0.1, and then performing a PCA on those, finally passing the components that represented 95% of the variance as input for the MLA. This might be explained by the fact that different relevant HRV metrics might reflect the same underlying information, and PCA helps to reduce this information into fewer features.

The method we propose could be deployed in real time in a NICU setting, updating the probability of late onset sepsis every half hour, based exclusively on the heart rate of the patient. Although this method would require a 48h observation period before the first prediction is made, in order to calibrate the system for the individual infant, this is not an impediment given that late onset sepsis is defined as sepsis occurring after the first 72h of life.

2.5 Conclusion

Based on our findings, we propose a method for LOS diagnosis in premature neonates using MLA based on HRV. Our first recommendation is to include visibility graph indexes, which are a novel method for HRV analysis, alongside the traditional metrics for HRV, to construct the feature set. Likewise, we recommend using a calibration period of 48 hours, proposing the median value over this time as the baseline for each individual infant, and then measuring the variation

in regards to this value. For training the MLA we recommend using the period of 42 hours before the beginning antibiotic treatment in the case of the infected population, and continuous periods of equal duration in the control population. For feature selection we recommend performing PCA over the features with p-value under 0.1 when comparing the measurements for sepsis versus non-sepsis population. Finally, for studies with a small population we recommend using logistic regression for making the predictions.

It is also worth mentioning that an article based on the study presented in this chapter was published in the IEEE Journal Biomedical And Health Informatics.

One drawback of this study, however, is that it was done on a small dataset, which might have been one of the factors that favored logistic regression models over more complex MLAs. It also required extensive feature selection and feature engineering to account for changes over time in the HRV characteristics on the infants.

As more infants were added to the database in later stages of the work for this thesis, we were able to overcome these disadvantages by testing more complex MLAs, such as recurrent neural networks. While these models require larger datasets to train on, they offer the advantage of possibly having better results and requiring less feature engineering, given that recurrent neural networks are well suited to capture time patterns in the data that are related to the evolutive process of infection onset. In the next chapter we dive into the details of the usage of these models in our context.

Appendices

2.A Construction of the Visibility Graphs and Calculation of their Indexes

To construct the visibility graph (VG) every data point of the RR time series is transformed into a node, and the connectivity between nodes is defined with the visibility criterion proposed by Lacasa et al. [25]. By this criterion, two data values of the time series (t_a, y_a) and (t_z, y_z) have visibility, and therefore are connected, if any other data point (t_i, y_i) placed between them, so that $t_a < t_i < t_z$ fulfill the following condition:

$$y_i < y_z + (y_a - y_z)\frac{t_z - t_i}{t_z - t_a}$$

The horizontal visibility graph (HVG) is a subset of the VG, in which the connectivity between nodes is defined by the criterion proposed by Luque et al. [26], by which (t_a, y_a) and (t_z, y_z) have visibility, and therefore are connected, if:

$$\forall t_i \in t_a, t_z : y_a > y_i \text{ and } y_z > y_i$$

We then calculated four indices from the VG and HVG thus obtained, in order to give a numerical characterization of their properties.

2.A.1 Mean Degree

The degree of a node is the number of connections (or edges) it has. The mean degree (MD) of the graph is then calculated as:

$$\text{MD} = \frac{1}{N}\sum_{n=1}^{N} d_n$$

Where N is the total number of nodes in the graph, and d_n is the degree of node n [27].

2.A.2 Cluster Coefficient

The cluster coefficient (C) index quantifies how connected the neighbours of a node are. The local cluster coefficient of node y_n, c_n, is given by:

$$c_n = \frac{\text{number of triangles connected to } y_n}{\text{number of connected triples centered on } y_n}$$

Where a triangle corresponds to three nodes that are connected to each other, and a connected triple is a set of three nodes which can be reached from each other. In other words, a connected triple is equivalent to a path formed by two edges, and in this case with node y_n as the central node.

Finally, the cluster coefficient C is calculated as the average of all the local cluster coefficients of all the nodes in the graph [27]:

$$\mathrm{C} = \frac{1}{N}\sum_{n=1}^{N} c_n$$

2.A.3 Transitivity

The transitivity (Tr) index also measures the density triangles in the graph, and is a global version of the cluster coefficient. It is calculated as:

$$\mathrm{Tr} = \frac{\text{3 x number of triangles in the graph}}{\text{number of connected triples in the graph}}$$

The factor three assures that $0 \leq C \leq 1$, given the fact that each triangle can be seen as three different connected triples, one with each of the data points as the central nodes [27].

2.A.4 Assortativity

Assortativity (r) is a correlation coefficient between the degrees of the nodes on opposite ends of an edge. It is calculated as:

$$\mathrm{r} = \frac{\frac{1}{N}\sum_{n=1}^{N} j_n k_n - [\frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(j_n + k_n)]^2}{\frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(j_n^2 + k_n^2) - [\frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(j_n + k_n)]^2}$$

Where j_n and k_n are the degrees of the nodes at each end of the n^{th} edge, and N is the total number of edges in the graph.

A network is assortative if the connected nodes have comparable degree ($r > 0$), otherwise the network is disassortative ($r < 0$) [28].

2.B Optimization of Hyperparameters

For each MLA we tested the same hyperparameters for every variation of the dataset, in each iteration of the of the LOOC method. As explained in section 2.2.5, we did this by implementing a grid search with an 8-fold cross validation split of the training dataset. However, for simplicity, in this section we will show the different hyperparameters tested for each of the four MLAs, and the combination of hyperparameters that was chosen most frequently (as percentage of the times

it was chosen in the LOOC procedure) as the best configuration for the MLA when predicting on the variation of the dataset that gave the best results in the window of 6 hours before t_0, and which was presented in Table 2.2.

Possible Combinations		Best Hyperparameters	
Number of Neighbors	Weights	Best Combination	Occurrence
[5, 7, 11, 15, 21, 25, 31]	[uniform, distance]	[5, distance]	47.1%

Table 2.B.1 – Best hyperparametrs for KNN when trained with the best performing dataset for the 6 hour evaluation window.

As presented in Table 2.B.1, for the KNN algorithm we adjusted two hyperparameters: the number of neighbours to be considered, and the weights with which they were considered. Uniform weight gives the same weight to all the points in the neighborhood, while distance weight gives a higher weight to neighbours that are closer. The combination chosen most often when training with the best variation of the dataset was with 5 neighbors, and weighted distance. This combination was observed in 47.1% of the LOOC iterations.

Possible Combinations			Best Hyperparameters	
Penalty	C	Solver	Best Combination	Occurrence
l1	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	[liblinear, saga]	[l2, 0.1, lbfgs]	97.1%
l2	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	[lbfgs, saga]		

Table 2.B.2 – Best hyperparametrs for LogR when trained with the best performing dataset for the 6 hour evaluation window.

Regarding LogR, we tested two different sets of combinations of hyperparameters, which are presented in the left rows of Table 2.B.2. The combinations were restricted in this manner due to the fact that the *liblinear* (*Library for Large Linear Classification*) solver does not support *l2* penalty, while the *lbfgs* solver (based on the *Limited-memory Broyden–Fletcher–Goldfarb–Shanno* algorithm) does not support *l1* penalty. For this algorithm the best combination, as shown, was observed in 97.1% of the LOOC iterations.

Possible Combinations			Best Hyperparameters	
Max. Depth	Min. Samples Split	Num. Estimators	Best Combination	Occurrence
[50, 100, None]	[2, 3]	[300, 600, 1000]	[50, 2, 600]	26.5%

Table 2.B.3 – Best hyperparametrs for RandF when trained with the best performing dataset for the 6 hour evaluation window.

The results for the RandF are presented in Table 2.B.3. For this MLA we adjusted three

different hyperparamters: the maximum depth of the tree (shown in column *Max. Depth*), the minimum number of samples required to split a node (*Min. Samples Split*), and the number of trees (*Num. Estimators*). In this case the best combination of hyperparametrs was a maximum depth of 50, with minimum sample split of 2, and 600 decision trees; even though this was the most frequently chosen as best combination, it occurred in only 26.5% of the LOOC iterations.

Possible Combinations				Best Hyperparameters	
Kernel	Gamma	C	Degree	Best Combination	Occurrence
rbf	[auto, scale]	[0.01, 0.1, 1, 10, 20, 50, 100]	-	[rbf, auto, 0.1]	83.3%
poly	[auto, scale]	[0.01, 0.1, 1, 10, 20, 50, 100]	[3, 5, 7, 9]		

Table 2.B.4 – Best hyperparametrs for SVM when trained with the best performing dataset for the 6 hour evaluation window.

For the SVM we adjusted the kernel, gamma, and C. In the case where the kernel was a polynomial function (*poly*), there was a fourth hyperparameter to consider, which was the degree of the polynomial. As shown in Table 2.B.4, in this case the most recurrent best combination was chosen in 83.3% of the iterations of our method.

Bibliography

[1] B. J. Stoll, N. Hansen, A. A. Fanaroff, L. L. Wright, W. A. Carlo, R. A. Ehrenkranz, J. A. Lemons, E. F. Donovan, J. E. T. Ann R. Stark, W. Oh., C. R. Bauer, S. B. Korones, S. Shankaran, A. R. Laptook, D. K. Stevenson, L.-A. Papile, and W. K. Poole, "Late-Onset Sepsis in Very Low Birth Weight Neonates: The Experience of the NICHD Neonatal Research Network," *Pediatrics*, vol. 110, no. 2, pp. 285–291, Aug. 2002.

[2] A. Kumar, D. Roberts, K. E. D. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg, D. Gurka, A. Kumar, and M. Cheang, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Critical Care Medicine*, vol. 34, no. 6, pp. 1589–1596, Jun. 2006.

[3] H. Nguyen, S. Corbett, R. Steele, J. Banta, R. Clark, S. Hayes, J. Edwards, T. Cho, and W. Wittlake, "Implementation of a bundle of quality indicators for the early management of severe sepsis and septic shock is associated with decreased mortality," *Critical Care Medicine*, vol. 35, no. 4, pp. 1105–1112, Apr. 2007.

[4] J. Alverdy and M. Krezalek, "Collapse of the Microbiome, Emergence of the Pathobiome, and the Immunopathology of Sepsis," *Critical Care Medicine*, vol. 45, no. 2, p. 337–347, Feb. 2017.

[5] R. Singh, L. Sripada, and R. Singh, "Side effects of antibiotics during bacterial infection: Mitochondria, the main target in host cell," *Critical Care Medicine*, vol. 16, pp. 50–54, May 2014.

[6] V. S. Kuppala, J. Meinzen-Derr, A. L. Morrow, and K. R. Schibler, "Prolonged Initial Empirical Antibiotic Treatment is Associated with Adverse Outcomes in Premature Infants," *The Journal of Pediatrics*, vol. 159, no. 5, pp. 720–725, Nov. 2011.

[7] World Health Organization. (2018) Antimicrobial resistance. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/antimicrobial-resistance

[8] A. Malik, C. P. S. Hui, R. A. Pennie, and H. Kirpalani, "Beyond the Complete Blood Cell Count and C-Reactive Protein," *Archives of Pediatrics & Adolescent Medicine*, vol. 157, no. 6, pp. 511–516, Jun. 2003.

[9] K. D.Fairchild and M. O'Shea, "Heart Rate Characteristics: Physiomarkers for Detection of Late-Onset Neonatal Sepsis," *Clinics in Perinatology*, vol. 37, no. 3, pp. 581–598, Sep. 2010.

[10] C. J. Chiew, N. Liu, T. Tagami, T. H. Wong, Z. X. Koh, and M. E. H. Ong, "Heart rate variability based machine learning models for risk prediction of suspected sepsis patients in the emergency department," *Medicine*, vol. 98, no. 6, p. e14197, Feb. 2019.

[11] J. Moorman, D. Lake, and M. Griffin, "Heart Rate Characteristics Monitoring for Neonatal Sepsis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 1, pp. 126–132, Jan. 2006.

[12] S. Ahmad, A. Tejuja, K. D. Newman, R. Zarychanski, and A. J. Seely, "Clinical review: A review and analysis of heart rate variability and the diagnosis and prognosis of infection," *Critical Care*, vol. 13, no. 6, p. 232, Nov. 2009.

[13] W.-L. Chen, J.-H. Chen, C.-C. Huang, C.-D. Kuo, C.-I. Huang, and L.-S. Lee, "Heart rate variability measures as predictors of in-hospital mortality in ED patients with sepsis," *The American Journal of Emergency Medicine*, vol. 26, no. 4, pp. 395–401, May 2008.

[14] N. Marwan, N. Wessel, H. Stepan, and J. Kurths, "Recurrence based complex network analysis of cardiovascular variability data to predict pre-eclampsia," in *Proceedings of Biosignal 2010*, Jul. 2010.

[15] X. Sun, Y. Zhao, and X. Xue, "Analyzing spatial characters of the ecg signal via complex network method," in *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, vol. 3, Oct. 2011, pp. 1650–1653.

[16] Z.-G. Shao, "Network analysis of human heartbeat dynamics," *Applied Physics Letters*, vol. 96, no. 7, p. 073703, 2010.

[17] X. Li and Z. Dong, "Detection and prediction of the onset of human ventricular fibrillation: An approach based on complex network theory," *Physical Review E*, vol. 84, p. 062901, Dec. 2011.

[18] T. Madl, "Network analysis of heart beat intervals using horizontal visibility graphs," *2016 Computing in Cardiology Conference (CinC)*, pp. 733–736, 2016.

[19] T. Nguyen Phuc Thu, A. I. Hernández, N. Costet, H. Patural, V. Pichot, G. Carrault, and A. Beuchée, "Improving methodology in heart rate variability analysis for the premature infants: Impact of the time length," *PLOS ONE*, vol. 14, no. 8, pp. 1–14, 08 2019.

[20] S. P. Shashikumar, Q. Li, G. D. Clifford, and S. Nemati, "Multiscale Network Representation of Physiological Time Series for Early Prediction of Sepsis," *Physiological Measurement*, vol. 38, no. 12, pp. 2235–2248, Nov. 2017.

[21] M. Doyen, D. Ge, A. Beuchée, G. Carrault, and A. I. Hernández, "Robust, real-time generic detector based on a multi-feature probabilistic method," *PLOS ONE*, vol. 14, no. 10, pp. 1–22, Oct. 2019.

[22] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Frontiers in Public Health*, vol. 5, p. 258, 2017.

[23] M. Bolanos, H. Nazeran, and E. Haltiwanger, "Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2006, pp. 4289–4294.

[24] A. Bauer, J. W. Kantelhardt, P. Barthel, R. Schneider, T. Mäkikallio, K. Ulm, K. Hnatkova, A. Schömig, H. Huikuri, A. Bunde, M. Malik, and G. Schmidt, "Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study," *The Lancet*, vol. 367, no. 9523, pp. 1674–1681, May 2006.

[25] L. Lacasa, B. Luque, J. Luque, and J. C. Nuño, "The visibility graph: A new method for estimating the Hurst exponent of fractional Brownian motion," *EPL (Europhysics Letters)*, vol. 86, no. 3, p. 30001, May 2009.

[26] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, "Horizontal visibility graphs: Exact results for random time series," *Physical Review E*, vol. 80, no. 4, p. 046103, Oct. 2009.

[27] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: A survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007.

[28] M. E. J. Newman, "Assortative Mixing in Networks," *Physical Review Letters*, vol. 89, no. 20, Oct. 2002.

[29] L. B. Mithal, R. Yogev, H. L. Palac, D. Kaminsky, I. Gur, and K. K. Mestan, "Vital signs analysis algorithm detects inflammatory response in premature infants with late onset sepsis and necrotizing enterocolitis," *Early Human Development*, vol. 117, pp. 83 – 89, Feb. 2018.

[30] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "proc: an open-source package for r and s+ to analyze and compare roc curves," *BMC Bioinformatics*, vol. 12, no. 1, p. 77, 2011.

[31] F. Harrell. (2019) Statistically efficient ways to quantify added predictive value of new measurements. [Online]. Available: https://www.fharrell.com/post/addvalue/

[32] S. Chen, L. Kang, Y. Lu, N. Wang, Y. Lu, B. Lo, and G.-Z. Yang, "Discriminative information added by wearable sensors for early screening - a case study on diabetic peripheral neuropathy," in *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 05 2019, pp. 1–4.

[33] V. Oliveira, W. von Rosenberg, P. Montaldo, T. Adjei, J. Mendoza, V. Shivamurthappa, D. Mandic, and S. Thayyil, "Early postnatal heart rate variability in healthy newborn infants." *Frontiers in Physiology*, vol. 10, p. 922, 2019.

[34] W. An and M. Liang, "Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises," *Neurocomputing*, vol. 110, pp. 101–110, 2013.

[35] J. S. Calvert, D. A. Price, U. K. Chettipally, C. W. Barton, M. D. Feldman, J. L. Hoffman, M. Jay, and R. Das, "A computational approach to early sepsis detection," *Computers in Biology and Medicine*, vol. 74, pp. 69 – 73, 2016.

[36] F. Lamping, T. Jack, N. Rübsamen, M. Sasse, P. Beerbaum, R. T. Mikolajczyk, M. Boehne, and A. Karch, "Development and validation of a diagnostic model for early differentiation of sepsis and non-infectious SIRS in critically ill children - a data-driven approach using machine-learning algorithms." *BMC Pediatrics*, vol. 18, no. 1, p. 112, Mar. 2018.

[37] S. Mani, A. Ozdas, C. Aliferis, H. A. Varol, Q. Chen, R. Carnevale, Y. Chen, J. Romano-Keeler, H. Nian, and J.-H. Weitkamp, "Medical decision support using machine learning for early detection of late-onset neonatal sepsis." *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 326–336, Mar. 2014.

[38] A. J. Masino, M. C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C. P. Bonafide, F. Balamuth, M. Schmatz, and R. W. Grundmeier, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," *PLOS ONE*, vol. 14, no. 2, pp. 1–23, 02 2019.

[39] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU." *Critical Care Medicine*, vol. 46, no. 4, pp. 547–553, Apr. 2018.

[40] D. W. Shimabukuro, C. W. Barton, M. D. Feldman, S. J. Mataraso, and R. Das, "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial," *BMJ Open Respiratory Research*, vol. 4, no. 1, 2017.

[41] R. Joshi, D. Kommers, L. Oosterwijk, L. Feijs, C. van Pul, and P. Andriessen, "Predicting neonatal sepsis using features of heart rate variability, respiratory characteristics, and ecg-derived estimates of infant motion," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 3, pp. 681–692, 2020.

[42] S. Ahmad, T. Ramsay, L. Huebsch, S. Flanagan, S. McDiarmid, I. Batkin, L. McIntyre, S. R. Sundaresan, D. E. Maziak, F. M. Shamji, P. Hebert, D. Fergusson, A. Tinmouth, and A. J. E. Seely, "Continuous Multi-Parameter Heart Rate Variability Analysis Heralds Onset of Sepsis in Adults," *PLOS ONE*, vol. 4, no. 8, pp. 1–10, Aug. 2009.

CHAPTER 3

Recurrent Neural Networks for Early Diagnosis of Late Onset Sepsis in Premature Infants Using Heart Rate Variability

3.1 Introduction

In Chapter 2 we discussed how changes in physiological signs, such as changes in heart rate variability (HRV), have been associated with neonatal late onset sepsis (LOS) [1], and detailed a study, carried out on a population of 49 premature infants, in which we showed that a simple machine learning algorithm, using only HRV features as input data, could detect LOS as early as 42 hours before the start of administration of antibiotics. In said study we showed the improvement of the models' performance when visibility graph indexes were included in the feature set, which allowed us to achieve an area under the receiver operating characteristics curve (AUROC) of 87.7% for the period of six hours before the start of antibiotics. However, that method required significant feature engineering, including a calibration period of 48 hours for each patient, to account for differences between the starting state of the patients and their end state.

This is an important consideration when working with preterm infants, given that their HRV characteristics undergo considerable changes during their stay in neonatal intensive care unit (NICU), as their autonomic nervous system continues to mature [2]. Moreover, the process of sepsis onset and its effect on HRV might also evolve over time.

Therefore, in this chapter we propose the use of recurrent neural networks (RNN) for the diagnosis of LOS in preterm infants based on HRV. RNNs are a type of artificial neural network specially well suited to analyse time series and exploit time-dependant patterns, because, as explained in Chapter 1, the recurrent connections in the network allow it to have memory. Thus, it would eliminate the need for intensive feature engineering or calibration periods in order to succeed at the task of early LOS detection in premature infants. However, it would still have the advantages of the method we previously proposed, of being non-invasive and continuously available in NICU settings, and yielding a probability of LOS in close to real-time.

For this study we used a population of 259 premature infants, from whom we acquired and processed the ECG signal to extract the HRV time series and features. Two models were constructed: one using the HRV time series as input, and the other using the same HRV features

as described in Chapter 2. The population was split into a training set, used to train the RNN models, and the test set, used to evaluate the performance of the models. The evaluation was done using the AUROC as the main metric. Finally, we also evaluate results from sample cases to analyse how the proposed models could be used as a decision support system (DSS).

3.2 Materials and Methods

3.2.1 Population

The data used in this study is part of the Digi-NewB cohort (NCT02863978, EU GA n°689260). The cohort prospectively included infants born between 25 and 42 weeks of gestation, hospitalized in the NICU of six university hospitals in France (University Hospitals of Rennes, Angers, Nantes, Brest, Poitiers, and Tours) in 2017-2019. Data collection was done with approval by the ethics committee (CPP Ouest 6-598) and informed parental consent.

For this study, we considered only the premature infants in the cohort, born before 30 weeks of gestational age; this included the infants in the population used for the study presented in Chapter 2 and additional infants who were added to the database after said study was carried out. Further selection of the patients was done retrospectively by a group of experts in neonatal medicine, who classified the infants into either LOS or control group. This classification was done according to the NEO-KISS protocol for nosocomial infection surveillance [3].

The resulting population consisted on 259 infants, of which 218 were in the control group, and 41 in the LOS group. Both groups were further split by randomly choosing 75% of the population of each group for the training set, and the remaining 25% for the test set. The final number of infants in each group is detailed in Table 3.2.1. We used only the infants in the train set to train the models, and reserved the test set only for the evaluation of the models.

Group	Control	LOS	Total
Train Set	163	30	193
Test set	55	11	66
Total	218	41	259

Table 3.2.1 – Population

3.2.2 Signal Processing and HRV Features Extraction

The electrocardiograms (ECGs) were obtained with a sampling rate of 500Hz. R-peak detection was done with a modified version of the Pan-Tompkins algorithm, with filter coefficients adapted for neonates [4]. This was done in the same manner as described in section 2.2.3. Afterwards, the R-R interval time series were extracted; this is referred to as the HRV time series.

At this point, two different segmentations of the HRV time series were done to generate two different feature sets.

To generate the first feature set, we split the time series into smaller segments with a fixed length of 1024 beats each; this is equivalent to approximately six minutes for an infant with a heart rate of 150 beats per minute. We retained the timestamp of the first beat in the segment as the timestamp associated with the segment. We refer to the feature set formed by these shorter HRV times series as the HRV_{1024} time series.

To generate the second feature set, the original HRV time series were split into segments with a fixed duration of five minutes. From each of the five minutes segments, we extracted the same HRV features detailed in Chapter 2, which are categorized in four different types: time-domain, frequency-domain, non-linear measurements [5], and visibility graph indexes. Finally, five minutes periods corresponding to 30 continuous minutes were grouped together by calculating the median value of each of their corresponding HRV features. This was done to minimize any noise in the data that could have resulted from artifacts in the ECG. Thus, the final feature set consists of the time series of 28 features that characterize the HRV, sampled in periods of 30 minutes. We refer to this time series as $\text{HRV}_{features}$.

3.2.3 Data labeling

Given that the objective of early diagnosis systems is to reduce delays in the beginning of treatment, we have defined the time of the LOS onset (denoted as t_0) as the time of beginning of administration of antibiotics.

For the training phase of the algorithm, both the HRV_{1024} and the $\text{HRV}_{features}$ datasets were labelled according to the following considerations. For the patients in the LOS group we included all available data before t_0. Based on the findings presented in Chapter 2, for the patients in the training set belonging to this group we labeled all the time segments within the period of 42 hours before t_0 as LOS, and any previous segments as not infected. For patients in the control group we included all available data up to the eighth day of life, so the average length of the time series for each patient in both groups would match. For the purpose of comparison with the LOS group when evaluating the performance of the machine learning models, we assigned the time of the last segment within this period as the t_0 for each patient in the control population.

It is important to note that the labeling for the evaluation of the model phase, particularly of the infants in the LOS group, was done differently. For the purpose of evaluating the behaviour of the performance of the machine learning models in the entire population in terms of the AUROC, the infants who belong to the LOS group were considered as always infected. We refer to this as the *population label*, as it was done on a population basis: patients who belong to the LOS population are labeled as LOS, and the patients in the control population as control. For the purpose of observing the behaviour of the predictions for a single patient, we labeled each

segment for the patients in the test set in the same manner as described for the training phase. We refer to this as the *segment label*, as it is done on a segment basis; for patients from the LOS group this means that only the segments belonging to the period of 42 hours before t_0 are labeled as LOS, while segments prior to this are labeled as control and, as usual, all segments for the control patients are labelled as control.

3.2.4 Recurrent Neural Network

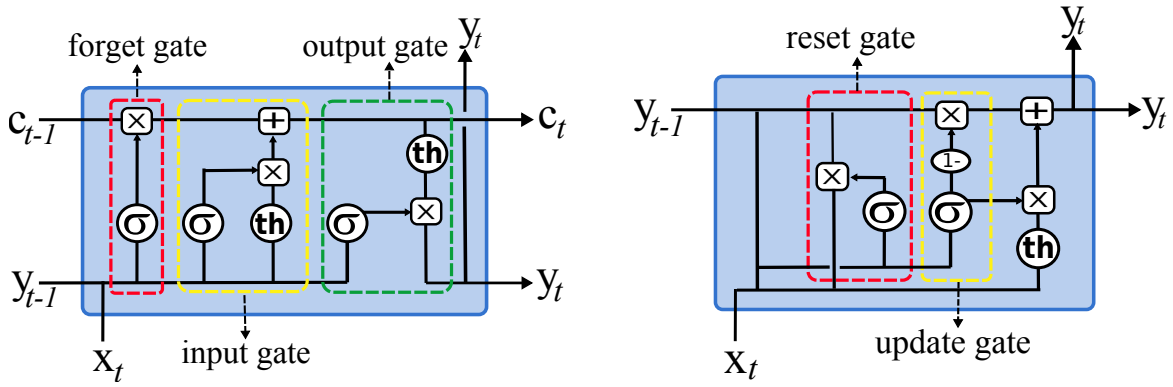
RNN are a type of artificial neural networks characterized by having recurrent connections [6]. In this type of architecture, each unit has a hidden recurrent state whose activation at each time step depends on the previous step. This feature allows RNNs to have memory, making them specially well suited to analyse time series and detect time-dependant patterns and changes.

Another characteristic of RNNs is that they can handle input sequences of variable length [6]. This makes RNNs a good choice for the problem of LOS diagnosis, given that the length of the input time series might vary for each infant in the dataset, based on when they developed LOS, or how much available data they have.

For the present study we used two types of RNNs units that are widely popular in recent literature: long short-term memory (LSTM) units and gated recurrent units (GRU). Both types of units have the capacity to capture dependencies of different time scales.

The characteristic feature of LSTM units [6, 7] is that they maintain a memory cell. The output of the LSTM unit is then a function of the current input and of the content of the memory cell. This is regulated by the output gate, which modulates the amount of memory content exposure at the output. Furthermore, LSTM units also have a forget gate, which modulates how much of the memory content is forgotten, while the input gate modulates how much new content is added to the memory cell. Thus, the memory cell can be updated by partially forgetting the existing memory and adding new content. Finally, both the output and the content of the memory cell are transmitted to the next time-step. This is the characteristic that allows LSTM units to easily carry relevant information over many time-steps in the input sequence, thus capturing long-distance dependencies.

This is exemplified in the diagram presented in Figure 3.2.1a, where we observe that the LSTM unit for a time-step t, receives the output from the previous time-step, y_{t-1}, the contents of the memory cell from the previous output, c_{t-1}, and the input for the current time-step, x_t. These three inputs then pass through the forget cell, denoted by the dashed red line in Figure 3.2.1a, which decides which of the previous contents of the memory cell to forget. Next is the input gate (yellow dashed line), which decides what new content to update in the memory cell. Finally is the output gate (green dashed line), which using the contents of the input, the previous output, and the updated memory cell, calculates the output of the current time-step. Both the updated contents of the memory cell c_t and the output y_t are passed to the next time-step.



(a) Long Short-Term Memory Unit.

(b) Gated Recurrent Unit.

Figure 3.2.1 – Diagram of the RNNs units. Adapted from [8] and [9]

GRUs [6, 10], unlike LSTM units, do not have memory cells. However, they are still able to capture time dependencies, as they also have gates that regulate the flow of information inside the unit. As shown in Figure 3.2.1b, the output of the GRU is a function of the output of the previous time-step and the input of the current one. This is modulated by a reset gate, enclosed in the red dashed lines in Figure 3.2.1b, which decides how much of the previous output to forget, and an update gate (yellow dashed line), which regulates how much of the unit's content is updated at each time-step.

Each gate within the unit, whether it is an LSTM or a GRU, is similar to a single neuron or unit in a feed-forward artificial neural network, in that it multiplies the inputs of the gate by a matrix of weights before adding them, and then applies an activation function to the result of the weighted sums. In this case, the activation function is either a hyperbolic tangent (represented by *th* in the Figure 3.2.1) or a sigmoid function (represented by σ) [11]. The most standard and widely used configuration of these units is using the activation functions as presented in Figure 3.2.1. The $\mathbf{x}$ and $+$ signs in the diagram represent pointwise product and pointwise sum, respectively. The dimensions of the output, which are the same as the dimensions of the memory cell in the case of LSTM, are determined by the dimension of the weight matrices in every gate. Therefore, while the number of columns of the weight matrices might vary, depending on if it is the weights associated with the input or with the output of the previous time-step, the number of rows must be the same for every matrix and it is also the size of the output. This is the dimension which is referenced to as the size of the cell or the size of the unit.

In RNNs, as in most machine learning algorithms, the weights are the parameters which the model needs to learn and adjust during the training phase. However, unlike classical feed-forward artificial neural networks, RNNs are not composed by multiple units in every layer. Instead, one layer uses the exact same unit (with the same weights) on each input. The output corresponding to the current time-step forms a feedback loop, connecting back to the same unit,

when it receives the input for the next-step. Thus, the RNNs architecture for each layer consists of exactly the same unit connected sequentially, hence the name *recurrent* neural networks. This is also the characteristic that allows the same model to adapt to sequences of different length as the input.

3.2.5 Proposed Models

In this study we aimed to test the performance of RNN models on two variants of the dataset. The first variant, as explained in Section 3.2.2, was constructed by taking the HRV_{1024} times series, thus using the raw HRV signal. The second variant was constructed by using the $\text{HRV}_{features}$ time series, thus using more processed features, but still without requiring the same level of extensive feature engineering and feature selection methods as used in the study presented in Chapter 2.

Based on preliminary tests in which we varied the architecture and hyperparameters of the RNN, we chose slightly different RNN architectures for each variant of the dataset. The different model variations that were tested are presented in Appendix 3.A, while in the following paragraphs we describe only the models that had the best performance, and on which the results presented in this Chapter are based.

RNN model for raw HRV times series

The model we implemented for LOS detection based on the raw HRV time series, HRV_{1024} consists of an input layer, three hidden layers, and an output layer. The input layer takes the raw HRV time series with length of 1024 beats, which is equivalent to one time-step in the input sequence. The first two hidden layers use LSTM units. Specifically, the first LSTM layer has internal cell size equal to 256, and the second LSTM has internal cell size equal to 64. The third layer consists of a fully connected layer, which is the typical feed-forward layer architecture presented in Section 1.1.6. This layer was constructed with 32 fully-connected units, using the rectified linear units (ReLU) [12] function as activation function. Finally, the output layer consists of one unit with a sigmoid activation function that returns the probability of LOS for each time-step as output.

RNN model for HRV featues time series

The model we implemented for LOS detection based on the time series of HRV features consists of an input layer, two hidden layers, and an output layer. The input layer takes the 28 HRV features from one time-step (equivalent to 30 minutes) at a time. The two hidden layers combine different types of RNN units, using a GRU, with internal sell size of 128, in the first hidden layer, and a LSTM unit, with internal cell size of 64 in the second layer. The third layer

consists of a fully-connected unit with a sigmoid activation function that returns the probability of LOS for each time-step as output.

Both models were optimized using binary cross-entropy as the loss function, which is a standard approach for binary classification problems, and the Adam algorithm as optimizer [13].

One important observation is that the data is strongly unbalanced, with more infants in the control than in the LOS group. Furthermore, for infants in the LOS groups, all time-steps belonging to the period previous to the 42 hours before t_0 are also labeled as control, further contributing to the data unbalance. Therefore, we assigned different weights to the samples of each class in the train set, according to the following rule:

$$\text{Sample Weight} = \begin{cases} 1 & \text{if the sample is labeled as control} \\ \dfrac{\text{Total number of control samples}}{\text{Total number of LOS samples}} & \text{if the sample is labeled as LOS} \end{cases} \tag{3.1}$$

Like so, the weighted sum of all the samples of one class will be equal to the weighted sum of the other, thus avoiding a bias of the loss function in favor of the majority class during training.

3.3 Results

In this section we first present the performance of the model for the HRV_{1024} dataset, followed by the results obtained with the $\text{HRV}_{features}$ dataset. Then we present use cases of the best performing model with infants from both the control and the LOS group, to demonstrate how this approach could be employed as a DSS in a NICU setting.

3.3.1 Predictive Performance of the Model Using the Raw HRV Time Series

Using the same evaluation methods as in Chapter 2, the proposed model for the HRV_{1024} dataset achieved an AUROC on the test set of 70.7%, with 95% confidence interval (CI) [68.6%, 72.9%], when evaluated on the six-hour window preceding t_0. In Figure 3.3.1 we show how the AUROC evaluated on the test set (shown in blue) changes on a sliding window of six hours, with a 3-hour overlap, during the 72 hours preceding t_0. We observe that the RNN model has an AUROC above 65% during the 18 hours before t_0. After that point, the AUROC begins to drop, reaching a minimum for the evaluation window of -36 to -42 before t_0. The CI does not display significant changes over the entire evaluation period.

However, we observe that the AUROC for the train set (shown in red) is consistently close to 100% for all the windows belonging to the period of 42 hours before t_0. This corresponds to the period that was labeled as LOS for the patients in the training set belonging to the LOS group.

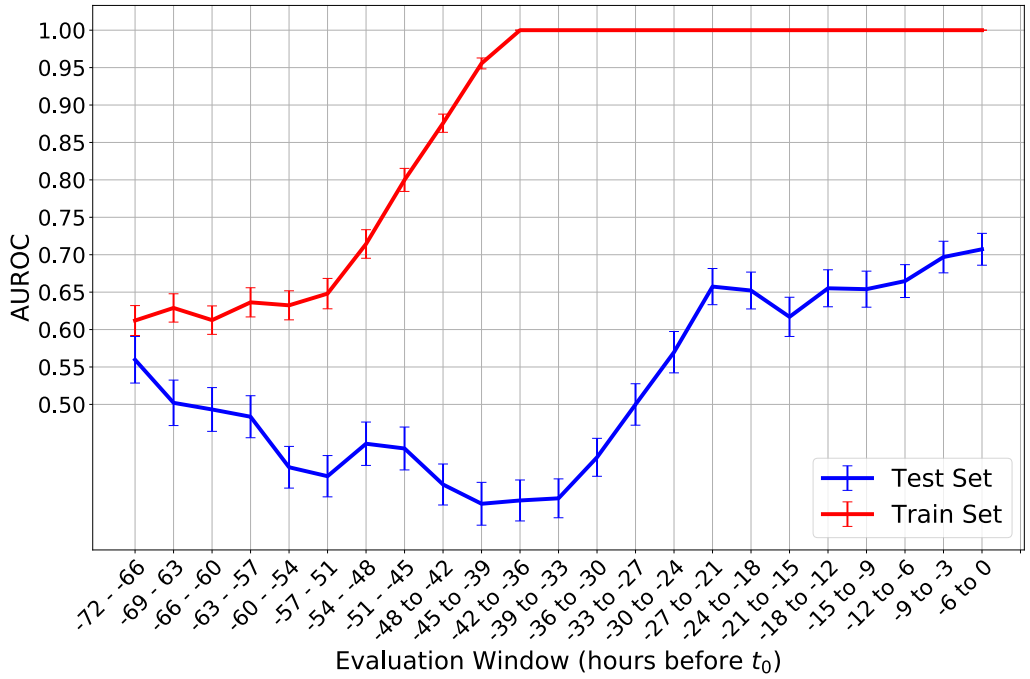


Figure 3.3.1 – Progress of the AUROC achieved by the RNN model on the raw HRV time series, evaluated on a sliding time window of six hours, with 50% overlap. In blue we present the results for the test set, and in red for the training set. The error bars represent the 95% CI.

After this point the AUROC drops dramatically, reaching nearly 60% for the last evaluation window, 72 hours before t_0. This is consistent with the fact that the model was trained using segment labels, so it only learned to detect sepsis for the 42 hours before t_0, and it is now being evaluated on the population label. Therefore classifications before the -42 hours as not infected for patients in the LOS group are penalized as miss classifications. The AUROC doesn't drop lower than 60% because the classification of control patients is the same with the segment labels as with the population labels, so the predictions for control infants are correct under this evaluation metric. From these observations we can infer that the model is strongly overfitting the training set.

3.3.2 Predictive Performance of the Model Using the HRV Features Time Series

Using the same evaluation methods, the model proposed for the $\text{HRV}_{features}$ dataset achieved an AUROC on the test set of 90.4%, with 95% confidence interval (CI) [88.1%, 92.6%], when evaluated on the six-hour window preceding t_0. In Figure 3.3.2 we show how the AUROC for the $\text{HRV}_{features}$ test set (shown in blue) changes on a sliding window of six hours, with a 3-hour overlap, during the 72 hours preceding t_0. We observe that the RNN model has an AUROC

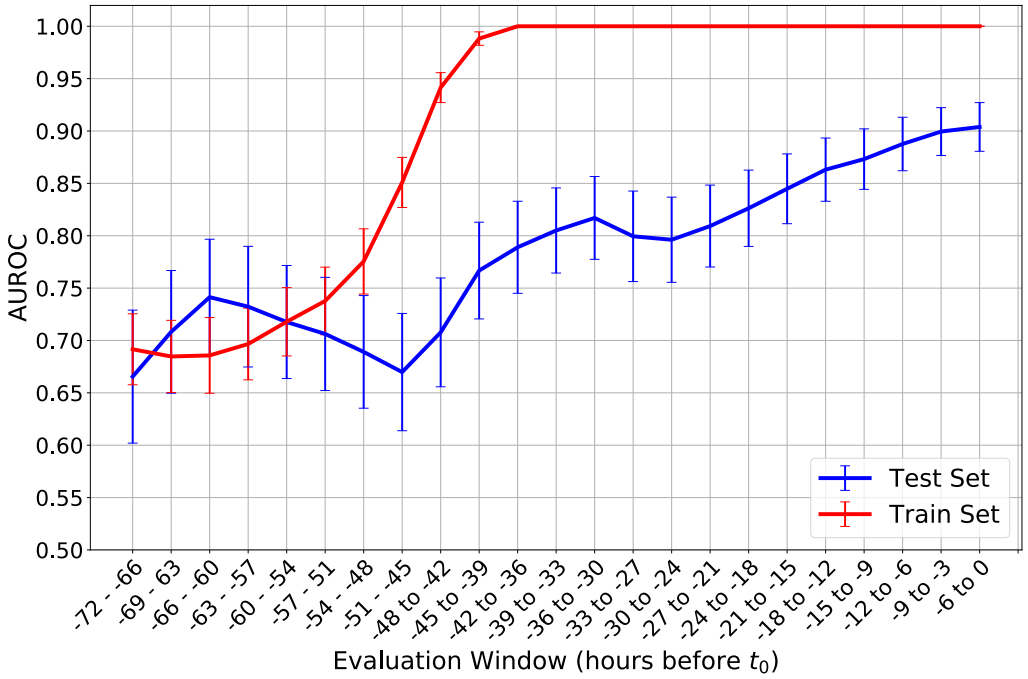


Figure 3.3.2 – Progress of the AUROC achieved by the RNN model on the HRV features time series, evaluated on a sliding time window of six hours, with 50% overlap. In blue we present the results for the test set, and in red for the training set. The error bars represent the 95% CI.

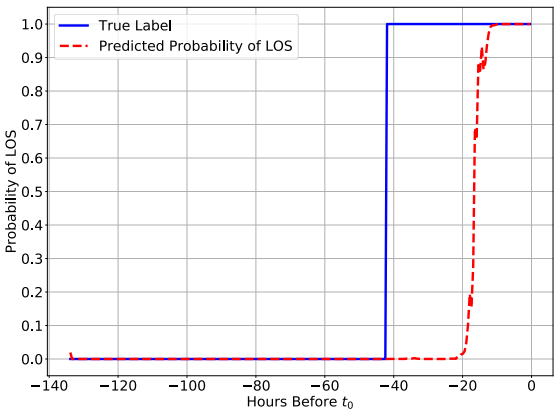
above 60% for all time windows. Furthermore, the AUROC is consistently above 70% since 48 hours before t_0, and above 80% for the 24 hours before t_0, until it peaks at 90.4% six hours before t_0. We also observe how the range of the CI is smaller for evaluation windows closer to the infection onset, at approximately $\pm 2\%$ for the 12 hours before t_0. It then gets progressively larger for evaluation windows further from t_0, peaking at 6.2% for the earliest evaluation window.

Similarly to the model for the HRV_{1024} dataset, we observe that the AUROC on the training set (shown in red) is also approximately 100% for during the 42 hours before t_0. This suggests that this model is also overfitting the data. However, the fact that it still manages a good performance, with AUROC of up to 90% on the test set indicates that the overfitting is less pronounced for this case and still manages to generalize reasonably well to data previously unseen by the model.

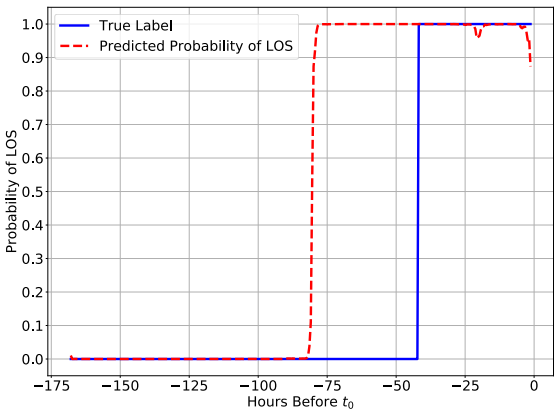
As this was the best performing model, in the next section we will observe in more detail its behaviour by examining the results for some individual patients.

3.3.3 Sample Cases

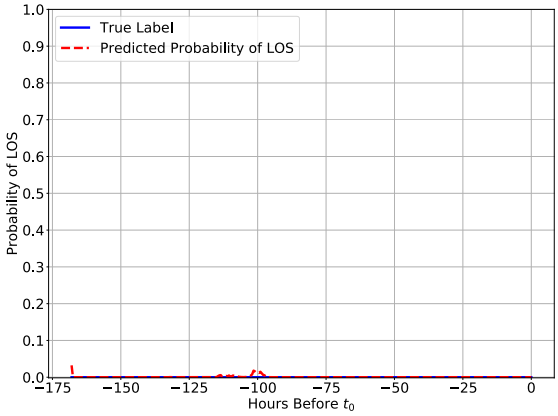
To exemplify how the proposed method could be used as a DSS, in Figure 3.3.3 we show the performance of the best model on three sample cases of patients belonging to test set. In



(a) LOS Group Patient 1



(b) LOS Group Patient 2



(c) Control Group Patient

Figure 3.3.3 – Examples of the predicted probabilities of a patient having LOS. The blue solid line represents the label assigned to a time period, and the red dashed line represents the probability of LOS calculated by the RNN model.

this section we focus only on the RNN model that was trained and tested on the $\text{HRV}_{features}$ dataset, due to the fact it significantly outperformed the model that used the HRV_{1024} dataset.

In Figure 3.3.3a, we present the results obtained for a patient from the LOS group (LOS group patient 1). As mentioned in Section 3.2.3, for the patients in this group we labeled the period of the 42 hours before t_0 as infected, and the rest as not infected. However, we observe that the model estimates a very low probability of LOS from the beginning of the studied period and until approximately 20 hours before t_0, at which time it starts detecting a very high probability of LOS for this patient; the probability remains very close to one until the end of the studied period.

In contrast, in Figure 3.3.3b we observe the results for another patient belonging to the LOS group (LOS group patient 2), for whom the probability of LOS estimated by the model

rises before the 42 hour period labeled as sepsis. In fact, for this patient the probability of LOS estimated by the model increases around 80 hours before t_0, after which it remains close to one for the reminder of the time.

Finally, in Figure 3.3.3c, we present a control patient for whom we observe that the probability of LOS estimated by the model is close to zero for the duration of the entire studied period.

3.4 Discussion

The main contribution of the work presented in this Chapter is the proposal of a machine learning model, based on RNN architecture, that uses only HRV data to produce reliable, early diagnosis of LOS in preterm infants. The best model was based on the same features described in Chapter 2, including the visibility graph features, and achieves an AUROC of 90.4% for the period of six hours before infection onset, using only HRV features as input. This type of architecture offers the advantage of being able to detect patterns that develop over time, without requiring extensive feature engineering or feature selection processes.

Previous studies have suggested the use of RNN models, with vital signs and other clinical data as input, for detecting sepsis in adults ([14, 15, 16, 17, 18, 19]). Studies targeted to sepsis diagnosis in neonates using artificial neural networks have mostly focused on convolutional neural networks also using multiple signals as input [20]. One study proposed the use of a convolutional neural network model with HRV features as input, which could potentially detect sepsis approximately 22 hours before clinical diagnosis [21] .

Therefore, the method presented in this chapter represents a novel approach to LOS diagnosis in preterm infants, and it has the advantage relying exclusively on HRV data, which can be automatically processed from heart rate monitoring readings that are done continuously and routinely in NICU. This also allows it to produce the LOS detection nearly in real time, being able to measure an updated probability of LOS every 30 minutes. These characteristics make the proposed method easy and practical to potentially implement in a NICU setting as a DSS, complementing other clinical signs.

An interesting aspect of the results we obtained is that the AUROC of the best model increases as the evaluation window gets closer to t_0, while its 95% CI reduces. This is consistent with the fact that the labeling of the samples, although based on the results of our previous study, is still arbitrary as it is impossible to determine exactly when a particular patient became infected, but as the evaluation gets closer to t_0, the certainty that a patient is infected increases. This is exemplified in the two sample cases of LOS patients we presented, where for patient 1 the detection of LOS by the model occurred approximately 20 hours before t_0, while for patient 2, it occurred approximately 80 hours before t_0. These differences between the results of individual

patients might reflect the differences in the time each patient was infected, rather than periods of incorrect detection.

Finally, regarding model built with the HRV_{1024} dataset, even though its performance was not very good, with a maximum AUROC of 70.7% (nearly 20% lower than the model with the HRV features time series), it is not be disregarded completely. As we observed in Figure 3.3.1, although the AUROC is relatively low for the test set, in the training set it is fairly high for the period of 42 hours before t_0, which is the period in which the algorithm was trained to detect LOS in infected patients. This suggests that the reason why the performance of the model is so low is due to overfitting on the training set. This is not surprising, given that for the RNN models, the number of parameters the algorithm has to learn depends on the number of input features. In this case, each time-step has 1024 input features, as every point in the time series is treated as a feature. Therefore, this model has considerably more parameters to learn than the model using HRV features, which only has 28 features as input, but the population size is the same for both cases.

However, the fact that the raw HRV time series model is able to fit so accurately to the training data signals that the model is indeed capable of learning the characteristics of the HRV time series that are associated with LOS, but that it would require more training data to be able to generalize well to previously unseen inputs. This is also not surprising, as complex artificial neural networks usually require thousands of data points to train on.

Thus, it would be recommendable to do further studies with RNN models using HRV time series, or even the ECG directly, on a larger population of infants, as we now know this type of model can extract the relevant information from this less processed data. And this type of model not only has the advantage of needing even less preprocessing of the raw data, but also they can work even closer to real-time. For instance, for an infant with an average heart rate of 150 beats per minute, which is normal for a very preterm infant [22], a time series of 1024 beats is equivalent to approximately six minutes. This means that a model such as the one we proposed for HRV time series could give an updated probability of LOS every six minutes, on average.

3.5 Conclusion

In this chapter we proposed a method for early LOS diagnosis in a preterm infants, based on recurrent neural networks.

We found evidence that suggests that such methods are capable of learning to differentiate between LOS and control patients even when using raw HRV data as input. However, due to the large amount of parameters the model needs to learn for this, further research with bigger databases need to be done in order to avoid overfitting and achieve better results in data

previously unseen by the model. Alternatively, other methods such as data resampling or data interpolation to augment the database could also be tested.

We were also able to propose a model with a very high performance that uses HRV features, including visibility graph indexes, for early LOS detection. This method achieves an AUROC of more than 80% for the 24 hours before the clinical diagnosis of LOS and beginning of antibiotic treatment, and it had a maximum AUROC of 90.4% for the period of six hours previous to LOS onset. We also observed the behaviour of the model's prediction in sample cases from individual patients taken from the test set, both from the LOS and the control group. This serves as a proof of concept that such an approach can potentially be deployed as a decision support system in NICUs where, compounded with other clinical observations and the expertise of the healthcare personnel, it could aid to achieve an earlier and accurate LOS diagnosis.

However, further studies should be done to evaluate the feasibility of applying such an approach in real-life. Specifically, the model should be validated on an entirely new database of infants, and adjusted if necessary. After this validation, we recommend the system to be tested on a clinical trial, to evaluate its impact in infants' mortality and length of hospital stay.

Finally, the methodology and results regarding the RNN model based on HRV features were accepted as a conference paper to Computing in Cardiology 2021. This conference will take place in September 2021 in Brno, Czech Republic.

Appendices

3.A Optimization of the Model Architecture and Hyperparameters

To optimize architecture and hyperparameters of the recurrent neural network (RNN) models we tested eight different models for both the HRV_{1024} feature set and the $\text{HRV}_{features}$ feature set. Each of the eight models had the same architecture for both feature sets, meaning that they had the same number of layers as well as the same type of unit per layer. However, given the considerable difference in size of the input for each feature set, the units in the HRV_{1024} models generally had a bigger internal cell size than those of the $\text{HRV}_{features}$ models. The types of units used to build the models were gated recurrent units (GRU), long short-term memory (LSTM) units, and fully connected (FC) units, which are also referred to in the literature as densely connected units.

In Table 3.A.1 we describe the eight models tested for each feature set. For each model we present the layers from bottom layer (closer to the input) to upper layer (closer to the output). The last layer, or output layer, of the models is always a fully connected layer with a single unit, so that the models' output is always a single number, which represents the probability of LOS for the time-step given as input. As mentioned before, in the Table we observe that the architecture, given by the type and number of layers, is the same for both HRV_{1024} and $\text{HRV}_{features}$ models, and only the cell size of the unit in each layer varies between the HRV_{1024} and $\text{HRV}_{features}$ for each of the tested model architectures.

As mentioned in Section 3.2.5, the best performing model, in terms of the highest area under the receiver operating characteristic curve (AUROC), for the HRV_{1024} input was different than the best performing model for the $\text{HRV}_{features}$ input.

— The best model for HRV_{1024} was model number *5*, with two consecutive LSTM layers, of 256 and 64 internal cell size, respectively, followed by two FC layers, of cell size 32 and 1, respectively. This model is presented in Figure 3.A.1.

— For the $\text{HRV}_{features}$ input the best model was number *4*, with a layer consisting of GRU with internal cell size of 128, followed by an LSTM unit layer of size 64, and finally the FC layer with a single unit. This model is presented in Figure 3.A.2.

Model	**Layers**	**Cell Size/Number of Units**	
		HRV$_{1024}$ Model	**HRV$_{features}$ Model**
1	LSTM	256	128
	LSTM	128	64
	LSTM	64	32
	FC	1	1
2	LSTM	256	128
	LSTM	64	64
	FC	1	1
3	LSTM	256	128
	GRU	64	64
	FC	1	1
4	GRU	256	128
	LSTM	64	64
	FC	1	1
5	LSTM	256	128
	LSTM	64	64
	FC	32	32
	FC	1	1
6	GRU	256	128
	GRU	128	64
	LSTM	64	32
	FC	1	1
7	GRU	128	64
	LSTM	64	32
	FC	1	1
8	GRU	256	128
	GRU	64	64
	FC	1	1

Table 3.A.1 – Architectures and Hyperparameters Tested to Optimize the RNN Models. The layers for each model are presented from bottom or input layer, to upper or output layer. In the case of the LSTM and GRU layers, the *Cell Size/Number of Units* columns refer to the size of the unit, while in the case of FC layers, the *Cell Size/Number of Units* columns refer to the number of units in the layer.

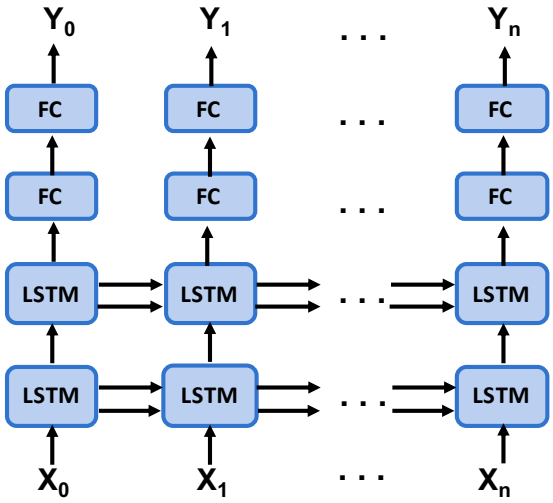


Figure 3.A.1 – Architecture of the best performing model for the HRV_{1024} input feature set, given by model *5* in the Table 3.A.1. X_0 represents the input corresponding to the first time-step for a given patient, in this case equivalent to 1024 consecutive beats, and Y_0 represents the corresponding output, which gives the probability of LOS for that time-step. X_n and Y_n represent the input and output, respectively, for the last time-step of the series corresponding to the patient.

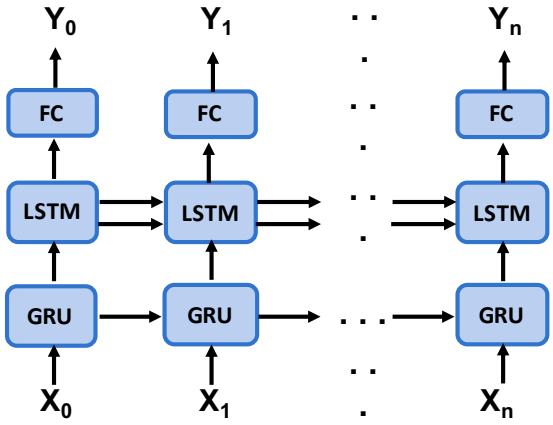


Figure 3.A.2 – Architecture of the best performing model for the $HRV_{features}$ input feature set, given by model *4* in the Table 3.A.1. X_0 represents the input corresponding to the first time-step for a given patient, in this case equivalent 30 consecutive minutes, and Y_0 represents the corresponding output, which gives the probability of LOS for that time-step. X_n and Y_n represent the input and output, respectively, for the last time-step of the series corresponding to the patient.

.

Bibliography

[1] K. D.Fairchild and M. O'Shea, "Heart rate characteristics: Physiomarkers for detection of late-onset neonatal sepsis," *Clinics in Perinatology*, vol. 37, no. 3, pp. 581–598, Sep. 2010.

[2] T. Nakamura, H. Horio, S. Miyashita, Y. Chiba, and S. Sato, "Identification of development and autonomic nerve activity from heart rate variability in preterm infants," *Bio Systems*, vol. 79, no. 1-3, pp. 117–124, 2005.

[3] National Reference Center For Nosocomial Infection Surveillance at the Institute for Hygiene and Environmental Medicine Charité – University Medicine Berlin, "NEO-KISS Protocol Nosocomial Infection Surveillance for preterm infants with birthweight $<$ 1500g," Feb. 2010.

[4] M. Doyen, D. Ge, A. Beuchée, G. Carrault, and A. I. Hernández, "Robust, real-time generic detector based on a multi-feature probabilistic method," *PLOS ONE*, vol. 14, no. 10, pp. 1–22, Oct. 2019.

[5] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, p. 258, 2017.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," Dec. 2014. [Online]. Available: https://arxiv.org/abs/1412.3555v1

[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997.

[8] C. Olah. (2015) Understanding LSTM networks. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[9] M. Phi. Illustrated guide to LSTM's and GRU's: A step by step explanation. Towards Data Science. [Online]. Available: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[10] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," Sep. 2014. [Online]. Available: https://arxiv.org/abs/1409.1259v2

[11] B. Ding, H. Qian, and J. Zhou, "Activation functions and their characteristics in deep neural networks," in *2018 Chinese Control And Decision Conference (CCDC)*, 2018, pp. 1836–1841.

[12] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," Feb. 2019. [Online]. Available: https://arxiv.org/abs/1803.08375

[13] J. B. Diederik P. Kingma, "Adam: A method for stochastic optimization," Jan. 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[14] H. J. Kam and H. Y. Kim, "Learning representations for the early detection of sepsis with deep neural networks," *Computers in Biology and Medicine*, vol. 89, pp. 248–255, 2017.

[15] T. Van Steenkiste, J. Ruyssinck, L. De Baets, J. Decruyenaere, F. De Turck, F. Ongenae, and T. Dhaene, "Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks," *Artificial Intelligence in Medicine*, vol. 97, pp. 38–43, Jun. 2019.

[16] M. Scherpf, F. Gräßer, H. Malberg, and S. Zaunseder, "Predicting sepsis with a recurrent neural network using the MIMIC III database," *Computers in Biology and Medicine*, vol. 113, p. 103395, Oct. 2019.

[17] Z. He, X. Chen, F. Zhen, W. Yi, C. Wang, L. Jiang, Z. Tong, Z. Bai, Y. Li, and Y. Pan, "Early sepsis prediction using ensemble learning with features extracted from LSTM recurrent neural network," 12 2019.

[18] C. Kok, V. Jahmunah, S. L. Oh, X. Zhou, R. Gururajan, X. Tao, K. H. Cheong, R. Gururajan, F. Molinari, and U. Acharya, "Automated prediction of sepsis using temporal convolutional network," *Computers in Biology and Medicine*, vol. 127, p. 103957, Dec. 2020.

[19] S. P. Shashikumar, C. S. Josef, A. Sharma, and S. Nemati, "Deepaise – an interpretable and recurrent neural survival model for early prediction of sepsis," *Artificial Intelligence in Medicine*, vol. 113, p. 102036, Mar. 2021.

[20] Y. Hu, V. C. Lee, and K. Tan, "An application of convolutional neural networks for the early detection of late-onset neonatal sepsis," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[21] P. Amiri, H. Abbasi, A. Derakhshan, B. Gharib, B. Nooralishahi, and M. Mirzaaghayan, "Potential prognostic markers in the heart rate variability features for early diagnosis of sepsis in the pediatric intensive care unit using convolutional neural network classifiers," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 5627–5630.

[22] C. J. Alonzo, V. P. Nagraj, J. V. Zschaebitz, D. E. Lake, J. R. Moorman, and M. C. Spaeder, "Heart rate ranges in premature neonates using high resolution physiologic data," *Journal*

of perinatology : official journal of the California Perinatal Association, vol. 38, no. 9, pp. 1242–1245, 09 2018.

CHAPTER 4

Evaluation of maturation in preterm infants through an ensemble machine learning algorithm using physiological signals

4.1 Introduction

In Chapters 2 and 3 we used heart rate variability (HRV) to evaluate the risk of late onset sepsis (LOS). However, HRV can also be an useful and non-invasive tool for evaluating the status of the autonomic nervous system (ANS) [1]. In neonates and infants, HRV measurements could be used to evaluate the process of maturation of their ANS ([2, 3]). HRV might also have diagnostic value for different clinical situations in neonates and infants which are linked to either congenital or acquired autonomic dysregulation. For instance, sudden infant death syndrome (SIDS) [4], neonatal seizures [5], and hypoxic-ischemic encephalopathy [6] have been associated with altered HRV.

Infants born prematurely are more susceptible to some of the aforementioned conditions when compared to term infants ([7, 8, 9, 10, 11, 12, 13]). Therefore, premature infants constitute a population of particular interest for the evaluation of their HRV as it is linked to the maturation of their ANS.

However, healthy preterm neonates present an altered HRV compared to that of infants born at term [14]. This translates into significant differences in time-domain [15], frequency-domain ([16, 17, 15]), and non-linear HRV measurements ([16, 17, 15]). Furthermore, significant differences have been found in the HRV of premature infants with different degrees of prematurity [16], with HRV measurements approaching normal values with increased gestational age (GA) [18]. Although the HRV parameters show an improvement with chronological age [19], these differences in HRV continue to prevail even when the postmesntrual age (PMA) of the preterm infants reach term equivalent age ([17, 20]), and for months afterwards [21]. While normative data for the HRV of full-term neonates has been proposed ([2, 3]), this is not the case for preterm newborns.

All these factors pose an obstacle for evaluating the maturation of preterm infants based on their HRV. Moreover, using the HRV of preterm infants as a potentially diagnostic tool for clinical conditions associated to autonomic dysregulation presents a greater difficulty, as even the HRV of healthy preterm infants at theoretical term will seem abnormal if compared to that

of healthy neonates born at full-term [22].

Previous studies have reported successfully using machine learning algorithms for the estimation of the maturity of preterm infants. As we discussed in Section 1.2.2, one study used functional magnetic resonance imaging of preterm infants at term equivalent age for predicting the GA of the infants, which the authors proposed as a surrogate measure of the brain maturity [23]. Another study used features derived from the electroencephalogram of preterm infants to predict the PMA, which the authors proposed as a surrogate measure for the brain maturation of the infants [24]. However, while the link between HRV, GA, and brain maturity (particularly of the ANS) has been largely reported in the literature, and the use of HRV and machine learning has been suggested to predict the prognosis of infants in the perinatal period [25], we did not find any previous study that used HRV and GA in combination with machine learning techniques, to evaluate the maturation process of infants during the period after birth.

Therefore, in this chapter we propose a method based for the use of a machine learning, with HRV measurements and GA as inputs, and a functional maturational age (FMA) as output.The FMA thus predicted, and its deviation from the PMA, which is measured clinically, could potentially help physicians evaluate the maturation of premature infants throughout their stay in the neonatal intensive care unit (NICU). This could also aid to the early detection of abnormalities in the neurological development of the infants as they manifest in the HRV.

As we have mentioned, this study is framed in the Digi-NewB project, which has collected not only HRV data from the neonates in its cohort, but also respiration, movement, bradycardia, cry, and sleep data. Therefore, and although the method proposed in this study was initially designed and developed for estimation of the maturation based on HRV features, we wanted the approach and resulting algorithm to be general enough that it could be used for FMA estimation using other types of data available in the project. Thus, we also tested the proposed method on respiration rate variability (RRV) and bradycardia features, which were available in the project at the time of this study. Similarly to HRV, RRV and bradycardia also show different patterns with increasing GA and PMA of the infants ([26, 27, 28]), as well with certain pathologies in neonates ([29, 30, 31]).

In the following sections, we describe the population used for the study, and the signal processing from acquisition of the electrocardiogram (ECG) to extraction of the HRV features. We describe the data analysis performed, as well as the machine learning algorithm we developed to evaluate the maturation of the infants, and we explain the method used to evaluate the resulting predictions. We also explain the steps taken to make the approach generic to different data types available in the Digi-NewB project. Finally, we present the HRV features that were used by the algorithm and the evaluation of the performance of the machine learning model using these features, as well as with the other datasets used to validate the generalization of our approach. Then we show the results for some sample cases. In the last section we discuss these

results and compare them with others reported in the literature.

4.2 Materials and Methods

4.2.1 Population

The data used in this study is also part of the database of the Digi-NewB cohort (NCT02863978, EU GA n°689260). The cohort prospectively included infants born between 25 and 42 weeks of gestation, hospitalized in the NICU of six university hospitals in western France (University Hospitals of Rennes, Angers, Nantes, Brest, Poitiers, and Tours) in 2017-2019. The collection of data was carried out after approval by the ethics committee (CPP Ouest 6-598) and informed parental consent.

The selection of the population of infants used for this study was done based on a three-step clinical evaluation performed by two senior neonatologists. First, they selected the newborns from the Digi-NewB cohort who did not present any of the following exclusion criteria: chest compression for resuscitation at birth; severe neurological lesions (grade 3 or 4 intraventricular haemorrhage, white matter lesions, hypoxic-ischemic encephalopathy); early onset sepsis; late onset sepsis; enterocolitis; severe malformations; and preterm infants with a birth weight lower than the 10th percentile for their GA. Second, both nenonatologists verified that, based on the clinical health reports, the pre-selected infants presented trajectories during the entire period of observation that could be considered normal for their GA. Third, both neonatologists met to verify and share their evaluation, and in case of doubt or disagreement in one case, that infant was excluded from the population.

The population was split into five categories, according to the GA of the infant: extreme preterm (EP), very preterm (VP), late preterm (LP), early term (ET), and full term (FT). The cut-off GA for each group, as well as the number n of infants in each group is detailed in Table 4.2.1. The definition of GA, PMA, and chronological age that we use throughout this chapter is that proposed by the American Academy of Pediatrics [32].

The HRV, RRV, and bradycardia population were built from the same base population of

Group	**GA (weeks)**	**HRV population (n = 50)**	**RRV population (n = 48)**	**Bradycardia population (n = 43)**
EP	[24, 28[	9	9	9
VP	[28, 32[	14	14	14
LP	[32, 37[	12	12	11
ET	[37, 39[	6	4	3
FT	≥ 39	9	9	6

Table 4.2.1 – Population characteristics

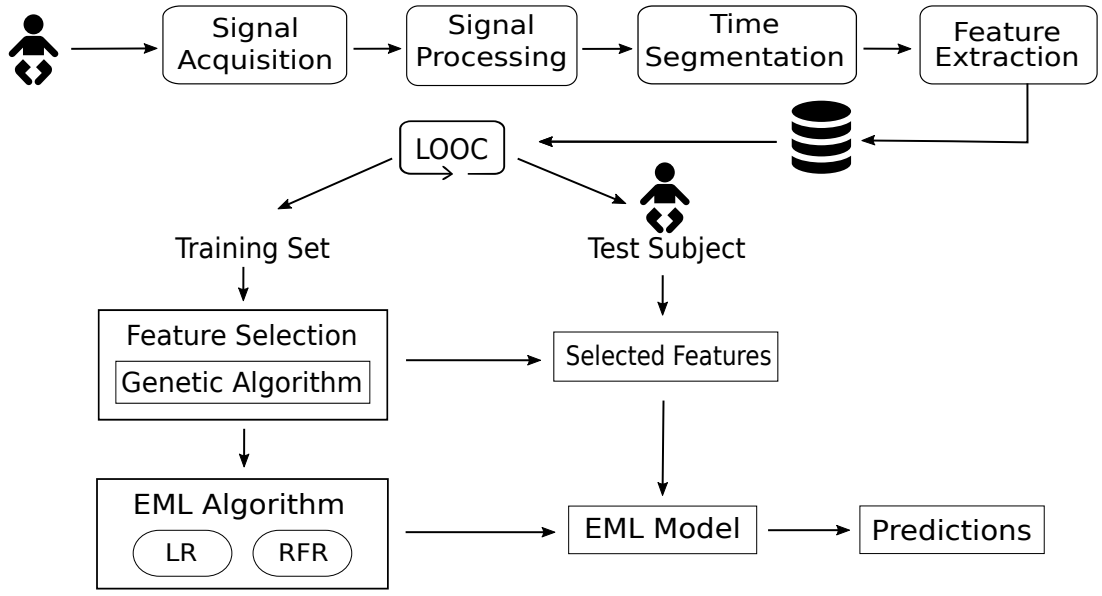


Figure 4.2.1 – Proposed approach.

50 infants, and the differences in the number of infants between them is due to the exclusion of neonates which did not have enough recordings of good quality in the case of the RRV population; or, in the case of the bradycardia population, because their data had not been processed yet for the detection of bradycardia episodes and the extraction of the associated features.

For preterm infants, the data was collected from continuous monitoring, during 24 hours a day, for the first three weeks of life. Afterwards, the monitoring was done also for 24 continuous hours, but every ten days. In the case of ET and FT infants, the data was collected only for 24 continuous hours after the third day of life, to avoid the phase of early adaptation to extrauterine life.

4.2.2 Proposed Approach

The general approach we propose is described in Figure 4.2.1. In general terms, we acquired and processed the raw signals from each patient, in this case the electrocardiogram (ECG) which was used for HRV, RRV, and bradycardia detection. The data is then segmented into shorter periods of time, which length might vary according to the data type. From these periods the features used to describe the HRV, RRV, or bradycardia are extracted.

Given the limited number of infants in our database, we used a subject based leave-one-out cross-validation (LOOC) technique. Using the LOOC technique can help minimize over-fitting of the machine learning algorithm, and the literature suggests it generally introduces less, or at most equal, variance and bias as using a K-fold technique [33]. Thus, all the subsequent steps of our algorithm iterate over the entire population, in each iteration one patient is used as the

test set (the patient left out), and the rest of the patients are used as the training set.

Next we used a genetic algorithm for feature selection and an ensemble machine learning (EML) algorithm [34] for the estimation of the PMA, which serves as the functional maturational age (FMA). The EML algorithm we propose combines the output of linear regression (LR) and random forest regression (RFR).

The following sections will explain in greater detail the different steps of our proposed approach. For signal processing and feature extraction we will detail only the HRV data, as this data type was the main focus of our study. Details concerning the signal processing to obtain the RRV data were proposed by Navarro et al. [35], and further information about the RRV features used in this study are given in the appendix 4.A. The signal processing for bradycardia detection was described by Altuve et al. [36], and information about the features used for this study are given in appendix 4.B. However, the subsequent steps regarding feature selection and the EML model are general and applicable to the different types of data.

4.2.3 Signal Processing and Extraction of the HRV Features

Using the same method described in Section 2.2.3, the ECGs were obtained with a sampling rate of 500Hz. R-peak detection was done with a modified version of the Pan and Tompkins algorithm, with filter coefficients adapted for neonates, as proposed in [37]. Afterwards, we extracted the R-R interval (RRI) time series, and then segmented it into 30 minutes periods, each of which was time-stamped so it could be associated to the corresponding PMA of the infant.

For each infant in the database, we selected only the available data corresponding to the first 12 weeks after birth. From each of the 30 minutes segments within the selected data we extracted the HRV features. These are the same features described in Section 2.2.4, which are categorized in four different types: time-domain ([38, 39, 39, 40]), frequency-domain [38], non-linear measurements [38], and visibility graph indexes ([41, 42]). These HRV features, along with the GA, compose our entire feature set and are shown in Table 4.2.2. We also retain the time-stamp to be able to calculate the true PMA of the infant associated to the features from each segment.

4.2.4 Data Analysis and Genetic Algorithm for Feature Selection

As a first step of our method for automatic feature selection, we calculated the Spearman correlation between all variables and the PMA in the training set, and eliminated all the features with a very weak absolute correlation to the PMA, ($|\rho_{xy}| < 0.1$). We did not eliminate any features based on their correlation to each other.

Next, we standardized the remaining features, as it is recommended to do when using linear regression models. Then we used a genetic algorithm [43] on the standardized features to find the

Category	Feature	Abbreviation
Time-Domain	Mean duration of the RRIs	meanRR
	Standard deviation of the RRIs	sdRR
	Root mean square of the RRIs	RMSSD
	Maximum RRI	maxRR
	Minimum RRI	minRR
	Skewness of the RRI time series	Skewness
	Kurtosis of the RRI time series	Kurtosis
	Acceleration of the heart rate	AC
	Deceleration of the heart rate	DC
Frequency-Domain	Low frequency power (0.02-0.2Hz)	LF
	High frequency power (0.2-2Hz)	HF
	LF in normalized units	LFnu
	HF in normalized units	HFnu
	The ration between LF and HF	LF_HF
Non-linear Measurements	Sample entropy	SampEn
	Approximate entropy	ApEn
	Short-range fractal correlation of the time series	$\alpha 1$
	Long-range fractal correlation of the time series	$\alpha 2$
	Short term variability derived from the Poincaré plot	SD1
	Long term variability derived from the Poincaré plot	SD2
Visibility Graph Indexes	Mean degree of the nodes in the VG	MD_VG
	Cluster coefficient of the VG	C_VG
	Transitivity of the VG	Tr_VG
	Assortativity of VG	r_VG
	Mean degree of the nodes in the HVG	MD_HVG
	Cluster coefficient derived from the HVG	C_HVG
	Transitivity of the HVG	Tr_HVG
	Assortativity of the HVG	r_HVG
Non-HRV related	Gestational age	GA

Table 4.2.2 – Category, description, and abbreviation of all features included in the feature set

optimal combination for our EML algorithm. The genetic algorithm was configured as follows:

Population We used a population of 20 chromosomes. In a genetic algorithm context, a chromosome encodes the information to be optimized. For this study, each chromosome had a length equal to the number of features, so each gene corresponded to one feature. The genes were binary, indicating if the corresponding feature was to be included or not in the final feature set. The first generation was initialized randomly, and subsequent generations were generated through a process of crossover and mutations.

Cost function To evaluate the cost (or fitness) of each chromosome, we used the mean absolute error (MAE) between the estimated FMA and the true PMA as cost function. For this, we trained the machine learning algorithm for which we were trying to optimize the features (either LR or RFR), on the feature set corresponding to each chromosome and calculated the resulting MAE of the FMA on the test set. The MAE was then used as the measure of cost, with a lower MAE indicating a lower cost. The algorithm stores the chromosome with the lowest cost as well as its associated cost. When a chromosome with an even lower cost is found, this information is updated.

Construction of new generations After the cost function was calculated for every chromosome of a given generation, a new generation would be constructed by ordering the chromosomes of the last generation from lowest to highest cost, and taking the top 30% (that is, the 30% best performing chromosomes) without changes and passing them on to the next generation. The remaining 70% of the new generation would be constructed by taking the top 50% of the last generation and designating them as the parent population; this population then undergoes a crossover and mutation process to generate their offspring which will pass on to the next generation.

Crossover The algorithm selects two chromosomes from the parent population at random, and performs a crossover operation, in which a new chromosome is created by taking half the genes from one parent, and the remaining half from the other.

Mutation The offspring chromosome resulting from a crossover operation goes through a mutation process, in which 10% of its genes are randomly chosen to change their value.

Stopping criteria The genetic algorithm would stop after reaching a maximum 150 generations or if the minimum cost remained constant for 30 continuous generations.

As the genetic algorithm is applied using the LOOC split of the population, it will result in one optimized set of features for each infant. Thus, to get one unique feature set to use in the EML algorithm, we retain only the features that appear in 50% or more of the optimized feature sets.

Since the ensemble model we propose uses both LR and RFR, we implemented two instances of the genetic algorithm: one with the target to minimize the MAE of the FMA (in weeks) obtained by a LR model, and one with the target to minimize the MAE of FMA estimated by the RFR model. Through this technique we obtained two new sets of features: one with the optimal features for LR and one for RFR, both optimized to minimize the MAE.

4.2.5 Ensemble Machine Learning

Based on preliminary inspections of the behaviour of the HRV features, we observed that some features displayed a linear relation to the PMA, while others seemed to have a non-linear behaviour. During the early stages of the study we also observed that LR and RFR favored different features. Therefore, we suggest the use of an EML model that combines LR and RFR. This allows the exploitation of both linear and non-linear correlations between the features derived from the physiological data and the PMA of the infants.

To build the EML model we first used the features selected as optimal for LR, by the method explained in the previous section, to train the LR part of the model. The FMA estimated by

this model, $\widehat{FMA}_{LR}$, is then added as an additional feature to the set of features selected as optimal for the RFR by the method described in section 4.2.4. Then the RFR is trained using this modified feature set as input, thus completing the training of EML model. Then, predictions are made on the test patient. When there are multiple observations (equivalent to 30-minute segments in the case of the HRV data) for the same day, the median value of all the FMAs corresponding to the same day and, therefore, to the same true PMA, is calculated and given as the predicted FMA for that day. Thus, the algorithm always gives only one FMA per day, regardless of the amount of available observations.

The same process is repeated for every iteration of the LOOC method, until every infant in the population has been used, at some point, as test patient.

4.2.6 Evaluation Method

To evaluate the accuracy of the model we compare the estimated FMAs with the PMAs of each infant in terms of the mean absolute error (MAE) in weeks, which is given by:

$$\text{MAE} = \frac{\sum_{n=1}^{N} |FMA_n - PMA_n|}{N} \tag{4.1}$$

where FMA_n is the age estimated by the model for a certain observation n, while PMA_n is the true PMA corresponding to the same observation. N is the total number of observations or predictions for each infant.

Given that we used a LOOC technique to train and test the model, so each infant in the population was at some point the test patient, we were able to calculate the MAE for each infant. Therefore, we then calculated the mean MAE, its 95% confidence interval (CI), the range of the MAE, and standard deviation (SD) over the entire population.

4.2.7 Generalization of the proposed method

In order to make the method described applicable to different types of data available in the project in which this study is framed, we had to make several considerations.

Handling missing values While data types such as HRV and RRV do not have any missing values in any time period for which recordings are available, this might not be the case for every type of data. For instance, in bradycardia data there might be missing values which correspond to features that would usually describe a type of bradycardia episode which did not occur during a period of observation. Thus, it was necessary to include a method that would allow to handle missing values.

The method incorporates different options for how to handle these cases. The user can choose to drop the features for which more than a certain percentage of observations are missing, with

the threshold also being adjustable by the user. To impute values on the missing data on the remaining features, the user can choose between using the mean, the median, or the most frequent value for that feature.

As a result, every feature F which has missing values, will be replaced by the feature F', which is constructed as follows:

$$F' = \begin{cases} f & \text{if the value was not missing in F} \\ f_i & \text{otherwise} \end{cases} \tag{4.2}$$

where f is the value in the original feature F, and f_i is the value calculated by the chosen imputation method.

Whichever imputation method is chosen, for every feature that has at least one missing value a new associated feature, *F_wasMissing*, is added to the dataset. This new feature is constructed as follow:

$$F\_wasMissing = \begin{cases} 1 & \text{if the value was missing in F} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

This feature is added because the information of whether a value was missing or not might be relevant [44]. This is the case for bradycardia, where the number of bradycardia episodes are expected to decline with increased PMA, so more missing values might be related to higher PMA.

Feature filtering by Spearman Correlation In this step we excluded the features with weak Spearman correlation to the PMA, before continuing to the genetic algorithm for feature selection. For this, we set the threshold at 0.1, so that features with an absolute correlation value under this threshold ($|\rho_{xy}| < 0.1$) would be eliminated from the feature set. We set such a low threshold because a combination of features that are weakly correlated to the target variable could contribute to the final prediction.

While datasets with many features, and many of which show a strong correlation with the PMA, might benefit from this filtering step, datasets with less features, or which have a weak correlation to the PMA, might not benefit from this. Thus, we have made this step optional.

Handling Categorical Features While HRV and RRV data only have numerical features, other data types might contain categorical features, or these categorical features might result from the missing data imputation process. Thus, we included the necessary provisions for the method to be able to handle both numerical and categorical features. Under these provisions, categorical features will be transformed into binary variables using one-hot encoding.

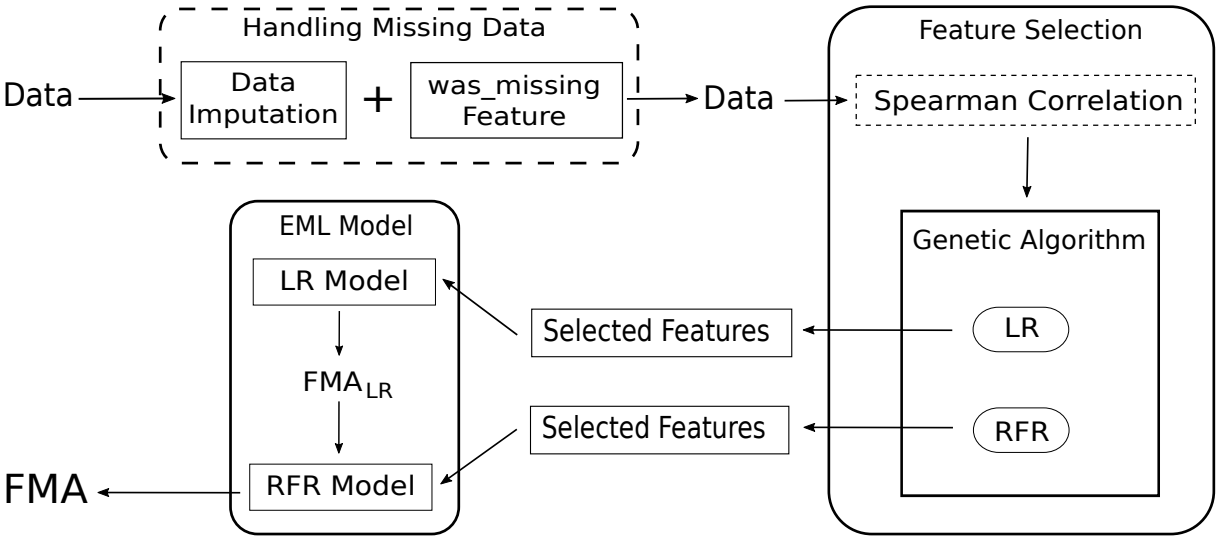


Figure 4.2.2 – Overview of the generic tool proposed. Optional phases and steps are represented framed by dashed lines. Phases and steps that are applied regardless of the data type are represented framed by solid lines.

The resulting generic tool is represented in Figure 4.2.2 and can be viewed as a series of blocks or phases. We represent phases or steps within each phase which are optional by framing them in dashed boxes; the phases or steps that are general, regardless of the data type used as input, are framed by solid lines. As observed in the figure, the method we propose takes the data, regardless of type, as input, and produces an estimation of the FMA as output. The first phase of the algorithm handles missing data by the method previously explained. This phase is optional as it applies only to data with missing values. The next phase is feature selection. While this phase is common to all data types, the step concerning filtering out the features with weak Spearman correlation to the target variables, as it was previously explained, is optional. The third phase is the EML model itself, where the FMA estimated by the LR model (FMA_{LR}) is used as an additional input feature by the RFR, thus producing the final estimated FMA.

All the data analysis, feature selection, EML algorithm, and evaluation process were developed in a Python environment.

4.3 Results

In this section we first focus on the results related to the HRV data, presenting the results of the feature selection method and the EML model for this data. Afterwards, we present the results obtained when applying the same method to RRV and bradycardia data. Finally, we show the results for two infants as sample cases.

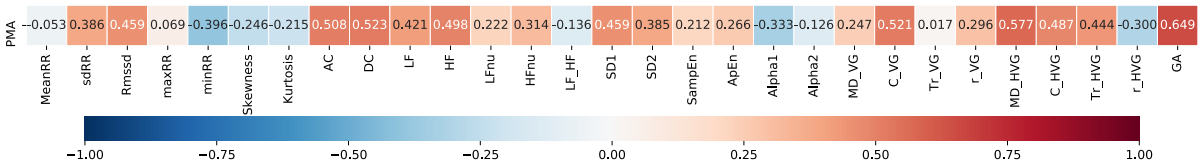
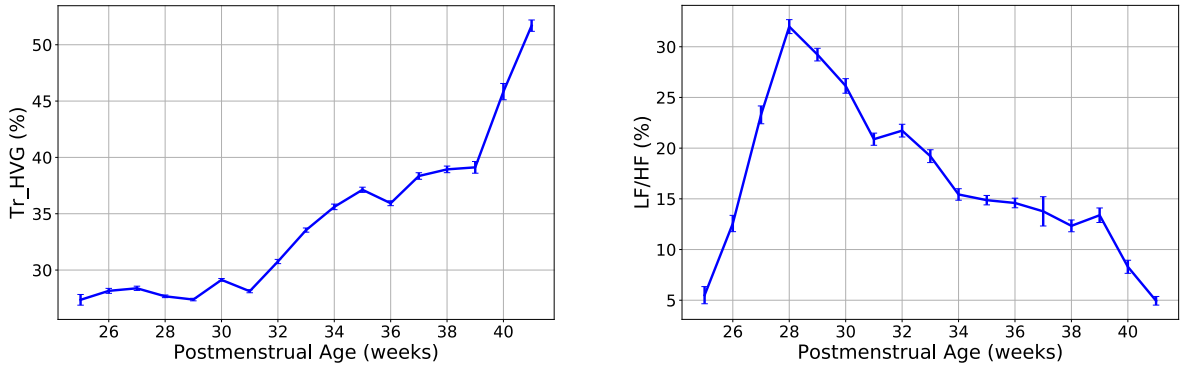


Figure 4.3.1 – Spearman correlation between HRV features and the target variable (PMA).



(a) Tr_HVG.

(b) LF_HF.

Figure 4.3.2 – Weekly average of the Tr_HVG (left) and the LF_HF (right) features for the entire HRV population. The error bars represent the standard error of the mean.

4.3.1 Selected HRV Features

The first step of the feature selection process was filtering out the features with a very weak Spearman correlation to the PMA ($|\rho_{xy}| < 0.1$). Figure 4.3.1 shows the correlations between all the HRV features, plus the gestational age, and the target variable (PMA). Based on this criteria, at this stage the only HRV features to be eliminated from the feature set were *meanRR*, *maxRR*, and *Tr_VG*.

The next step was using two instances of the genetic algorithm, each to optimize the feature set for training a LR model and a RFR model, respectively. This was done using the LOOC method, so we only retained for the final model the features that were present in at least 50% of the LOOC iterations. The features obtained by this technique are listed in Table 4.3.1, and in parenthesis we present the percentage of the LOOC iterations for which the feature was chosen. We observe that the LR model favors mostly the time-domain features and visibility graph indexes derived from the HRV, while the RFR model relies mostly on frequency-domain features. From the non-linear measurements, both models use only one feature (ApEn), and also both models use the GA, which is the only feature in the set which is not derived from the HRV.

The use of different features by the two different models indicates that some HRV characteristics might have a linear correlation to the PMA (mostly time-domain features and visibility

Category	Linear Regression	Random Forest Regression
Time-domain	sdRR (50%) minRR (70%) Skeweness (84%) Kurtosis (56%)	minRR (92%)
Frequency-domain	LFnu (62%) HFnu (58%)	LF (74%) HF (56%) LFnu (100%) HFnu (56%) LF_HF (76%)
Non-linear measurements	ApEn (60%)	ApEn (78%)
Visibility Graph Indexes	MD_VG (72%) C_VG 56%) r_VG (50%) C_HVG (80%) Tr_HVG (76%)	C_HVG (88%)
Non-HRV related	GA (90%)	GA (92%)

Table 4.3.1 – List of features selected by the genetic algorithm for the linear regression and random forest regression models. The percentage of LOOC iterations for which the feature was chosen is presented in parenthesis.

graph indexes), while others (mostly the frequency-domain features) might have a non-linear correlation to the PMA. This is exemplified in Figure 4.3.2, where we show the weekly average of two features for the entire HRV population. One of the features selected by the genetic algorithm exclusively for the LR model was *Tr_HVG*, shown in Figure 4.3.2a, for which we observe that its behaviour is, in general, monotonically related to the PMA. While Figure 4.3.2b we display the behaviour of the *LF_HF* feature, one of the features which was selected by the genetic algorithm exclusively for the RFR model and which does not show a linear relation to the PMA.

4.3.2 Performance of the EML model on HRV data

Data type	HRV	RRV	Bradycardia
Mean MAE (weeks)	0.93	1.39	1.39
Max. MAE (weeks)	2.01	4.33	4.31
Min. MAE (weeks)	0.03	0.02	0.03
SD (weeks)	0.54	0.93	0.99
95% CI (weeks)	[0.78, 1.08]	[1.12, 1.66]	[1.08, 1.69]

Table 4.3.2 – Performance of the model on the HRV, RRV, and bradycardia data, as measured by the mean, maximum, and minimum MAE, its SD and 95% CI.

We evaluated the accuracy of the EML model to estimate the functional maturational age (FMA) on the test patient in every iteration of the LOOC process. To this end we calculated

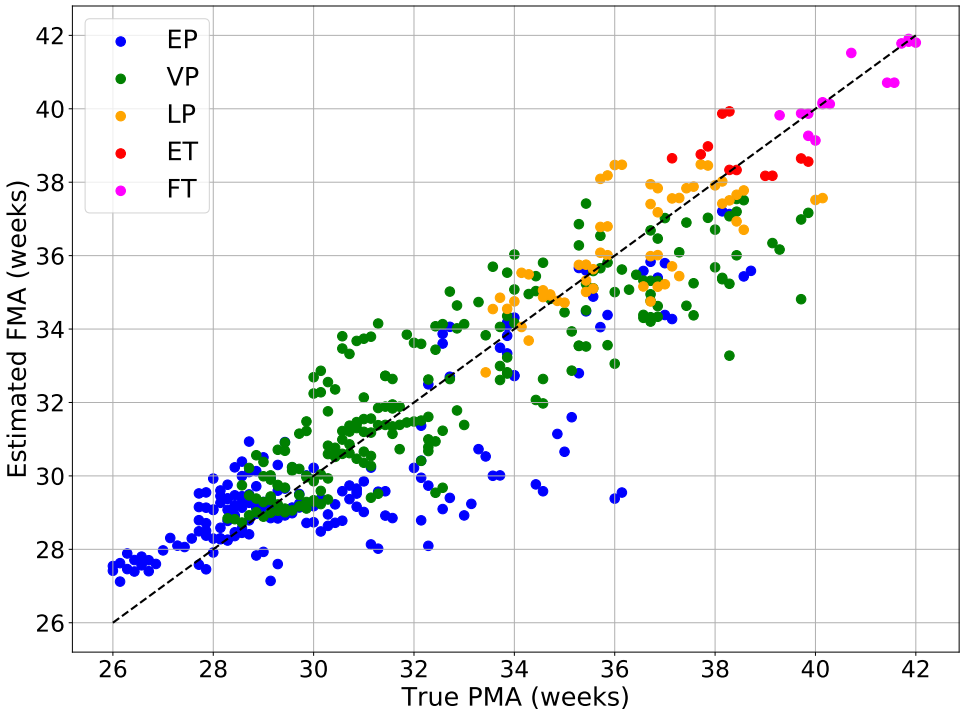


Figure 4.3.3 – PMA versus FMA for all infants in the HRV population, grouped by term.

the MAE, in weeks, of the FMA estimated by the model in relation to the PMA, for every patient. From this we obtained the mean MAE, maximum MAE (Max. MAE), minimum MAE (Min. MAE), the standard deviation (SD), and 95% confidence interval (95% CI) for the entire population. These results are summarized in the first column of Table 4.3.2. We observe that the mean MAE over the population is under one week, with the maximum at just over two weeks. This, accompanied by a low SD and a narrow 95% CI suggests that the method is robust for estimating the FMA.

Furthermore, in Figure 4.3.3 we observe the scatter plot of the PMA versus the FMA for the entire population, grouped by terms. We observe that most observations fall reasonably close to the dotted line which represents a perfect prediction, for which the PMA and the FMA would be equal.

Most of the observations that are far from this line belong to EP infants (blue). This is also the case for some VP infants (green). While LP (yellow), ET (red), and FT infants (magenta) tend to have less error.

4.3.3 Validation of the Model on RRV and bradycardia Data

We also evaluated the performance of the EML model when it was trained and tested on the bradycardia and the RRV data, respectively. The list of the selected features for these data

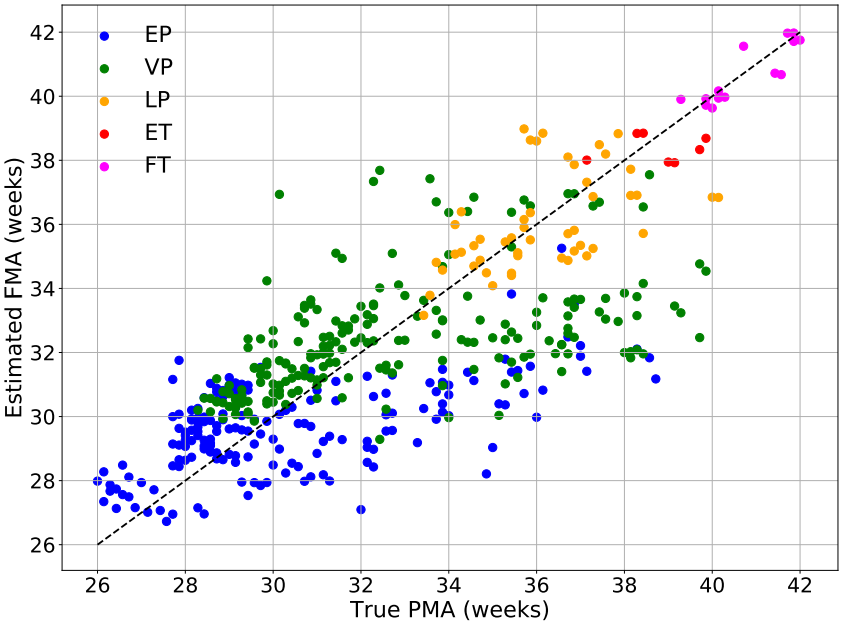


Figure 4.3.4 – PMA versus FMA for all infants in the RRV population, grouped by term.

types can be found in the Appendix 4.B and Appendix 4.A, respectively. The results regarding the estimation of the FMA are summarized in Table 4.3.2. We observe that both data types have very similar results. The mean MAE is low, at under 1.4 weeks for both cases. However, the performance is not as good as the one obtained on the HRV data, where we had a mean MAE of 0.93 weeks. Likewise, the Max. MAE is also higher for these data types, with a Max. MAE at around 4.3 weeks. On the other hand, the Min. MAE remains quite low and the result is comparable to that obtained with the HRV data. The SD for these data remains under one week, and the 95% CI also remains reasonably narrow. These results suggest that the method we propose for FMA estimation can be used on different data types with reliable results.

The higher error obtained when using the RRV and bradycardia data might be explained by the fact that these data types have, in general, a weaker correlation to the PMA as compared to the HRV data. This is particularly true in the case of RRV data, for which we opted to not perform the feature filtering by Spearman correlation, as it reduced the feature set to only a few variables. Another factor which might contribute to a lower performance is that both the RRV and bradycardia population (n=48 and n=43, respectively) were smaller than the HRV population (n=50), which translates into fewer data to train the model on, a difference that is specially marked for the bradycardia population, which is 14% smaller than the HRV population.

In Figure 4.3.4 we present the PMA versus FMA scatter plot for all the infants in the RRV population, while in Figure 4.3.5 we present the results from the bradycardia population. Although the results are more dispersed than for the HRV data, the general behaviour is the

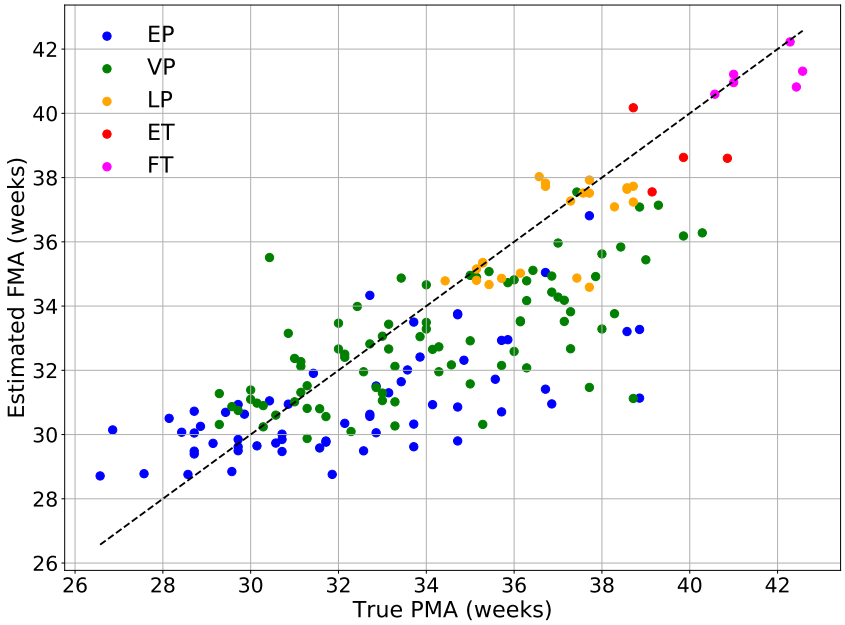


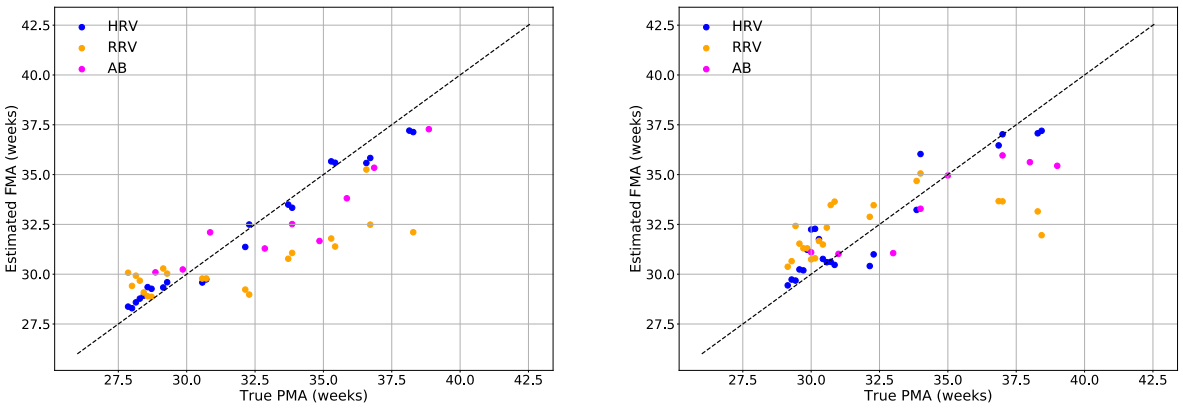
Figure 4.3.5 – PMA versus FMA for all infants in the Bradycardia population, grouped by term.

same, with EP (in blue) and VP (in green) infants presenting a smaller deviation in the FMA for smaller PMAs, and the deviation increasing as they approach the term equivalent age.

4.3.4 Sample Cases

In Figure 4.3.6 we display the scatter plot of the true PMA versus the FMA estimated by our model (in weeks) for two sample cases.

On the left (Figure 4.3.6a) we present the results obtained, with all three different types of data, for an EP female born by C-section at 27 weeks and 6 days (27.86 weeks) of GA. She was antenatally treated with corticosteroids and magnesium sulfate as recommended. Her birth weight was 930 g with good adaptation to extra-uterine life (Apgar scores: 8-10) and no particular complications during her hospital stay. We observe that for this baby, as the true PMA increases so does the FMA estimated by our model. While this is true for all three data types used, we observe that the FMA estimated using HRV data (presented in blue) are always closer to the dotted line which represents a perfect prediction (for which the FMA and the PMA would be equal). The results obtained with the bradycardia data (shown in magenta) are also relatively close to this line, while it is the RRV data (presented in yellow) which displays the greater dispersion and strays further from the dotted line. However, it is worth noting that for all the data types the FMA is usually below the PMA for the observations that correspond to greater PMAs. This is consistent with the fact that preterm infants display different characteristics than full-term infants, even when they reach term-equivalent age [17].



(a) EP infant.

(b) VP infant.

Figure 4.3.6 – True PMA versus Estimated FMA for sample cases, using HRV (in blue), RRV (yellow), and bradycardia (magenta) data.

On the right figure (Figure 4.3.6b) we presents the results obtained for a VP female born by C-scetion at 29 weeks of GA. She was antenatally treated with corticosteroids and magnesium sulfate as recommended. Her birth weight was 1045 g with good adaptation to extra-uterine life (Apgar scores: 2-10) and no particular complications during her hospital stay. Similar to the previous case, we observe in this figure that as the true PMA increases, so does the FMA estimated by our EML model with each data type. In this case both the HRV (blue) and bradycardia (magenta) estimations stay closer to the line which represents the perfect prediction, while once again it is the estimation obtained with the RRV data which shows the biggest error, especially for the observations surrounding the 37.5 weeks of true PMA.

4.4 Discussion

In the study presented in this chapter we developed an automated estimator of the FMA for preterm and full-term infants, which can be used as a surrogate measure of the maturational age. The estimator is based on an EML model which is capable of making estimations using different types of data as input, which could be extracted from the heart rate monitoring of the infants, which is typically done in a NICU. While we focused our study on using HRV data to estimate the FMA, we also showed that the model we propose can make estimations using RRV and bradycardia data.

Previous studies took similar approaches, using brain imaging [23] or electroencephalogram recordings of preterm infants [24] in combination with machine learning algorithms to estimate a functional maturational age of the infants. However, we did not find any previous study that used HRV data to produce such an estimation of the FMA. Thus, this study proposes a novel and non-

invasive approach to the evaluation of the maturation of infants using HRV data. Furthermore, all the previous studies we found that use machine learning to estimate a maturational age of infants were dedicated to a specific data type. We did not find any previous studies that propose a generic method for estimation of the maturational age using machine learning that can be applied to different types of data acquired in a NICU setting.

Another novel feature of this study is that it includes visibility graph indexes to characterize the HRV. This type network-based analysis has begun to be applied to HRV in the last decade [45], and its use for the analysis of the HRV of premature infants has only recently been suggested [46]. However, we showed in Chapter 2 that the inclusion of the visibility graph features add new information about the HRV characteristics, resulting in better performance of machine learning algorithms.

The EML model we propose combines LR and RFR to make the final predictions. This allows the model to exploit information from the features in the data, whether their correlation with PMA is linear or non-linear. To select the optimal feature set to train the EML model we used genetic algorithms. The feature selection process was done separately for the LR and for the RFR portions of the EML model, so that each element of the EML model would only use the features optimal for it. In the case of HRV, this yielded interesting results, showing that while LR predominantly favored time-domain and visibility graph features, RFR favored frequency-domains features, suggesting that the later have a non-linear correlation to the PMA of the infants.

An important characteristic of the method proposed in this chapter is that in the population we used to train and validate the model we included preterm infants with varying degrees of prematurity, as well as full-term infants, none of which manifested clinical signals of abnormal maturation for their GA. Consequently, this method is well suited to evaluate infants regardless of their GA. This is specially important in the case of the preterm infants because, to our knowledge, no normative HRV data has been proposed for this population.

The performance of the model was very accurate, with a mean MAE under one week when using the HRV data, and under 1.4 weeks when using the RRV and bradycardia data. Part of the error in the estimations done by our model might be accounted for by the inherent error associated to the PMA, due to the difference between the day of conception and the day of the last menstrual period.

However, another interesting observation was that EP and VP infants presented more deviation in their FMA when compared to their PMA. This might be explained by the fact that, as mentioned before, there are differences in the HRV of preterm infants at term equivalent age and that of infants born at term ([17, 20]). This explanation is supported by the fact that, for both the EP and the VP groups, the observations with the biggest difference between the PMA and the FMA occur for PMAs that are closer to term equivalent age. Thus, the increased error

as the preterm infants reach term equivalent age reflects this expected delay in the maturation of the preterm infants and suggests that the models we propose offers a good measure of the maturational age of the infants.

This suggests that the model we propose might be a reliable tool for estimating the FMA of neonates based on different physiological signals, and use it as surrogate measure of their maturational age. In a NICU setting, this estimated FMA and its deviation from the PMA could help physicians evaluate the maturation process of the infants in real time and without need for invasive or additional tests.

4.5 Conclusion

In this chapter we proposed an automated, non-invasive method for estimation of the maturation of infants during their first months of life, based on machine learning using different HRV, RRV or bradycardia features. As these features are extracted from the heart rate or respiration rate monitoring of the patients, this method has the potential to be used in real time and as a bed-side tool in NICU settings, given that the EML model would require only a minimum of 30 minutes of continuous heart rate or respiration rate monitoring to produce an estimation of the functional maturational age.

The method we proposed uses genetic algorithms to find the optimal features for the machine learning algorithm, and combines linear regression and random forest regression to estimate the FMA of the infant. We propose this estimated FMA as a surrogate measure of the maturation of the infants. This measurement and its possible deviations from the PMA, which is measured clinically, could assist clinicians in making decisions regarding assisted ventilation, discharge, sleep management, and environmental care of the infants.

The study presented in this chapter was also the subject of an article which was recently accepted with minor revisions in the IEEE Journal of Biomedical Informatics. It also led to the development of a software which is currently being licensed.

This chapter focused on the evolution of healthy infants during their hospitalization in NICU. In the next, we will test this method on a population of infants who presented complications during their hospitalization in NICU. This would allow us to test the hypothesis that for these infants the FMA estimated by the EML model should differ even more from the PMA than it does for the population of healthy infants presented here.

Appendices

4.A RRV Features

For extracting the features used to describe the RRV, the respiration signal was acquired from clinically applied ECG leads using the trans-thoracic impedance. The signal thus obtained was first segmented into 30-minutes periods. From each segment, three time series were computed:

— *Inspiration phase (T_{in})*: This time series reflects the variability in the duration of the inspiration phase, and was computed as the time differences between consecutive minima and maxima in the respiration signal.

— *Expiration phase (T_{ex})*: This time series reflects the variability in the duration of the expiration phase, and was computed as the time differences between consecutive maxima and minima in the respiration signal.

— *Breathing cycle (T_{tot})*: This time series reflects the variability in the duration of the entire breathing cycle, and was computed as the time differences between consecutive minima in the respiration signal.

From each of the these time series different features were extracted, resulting also in three category of features. A fourth category of RRV features was used to quantify the number of episodes of apnea.

The category, name, and a brief description of the features is presented in Table 4.A.1. For simplicity, we only present the features that were selected by the genetic algorithm for either the LR or RFR model, or both. The last two columns of the table represent whether the feature was included or not in the LR and/or RFR feature sets.

Feature Category	Feature	Feature Description	LR	RFR
$\mathbf{T}_{tot}$	Mean_T_{tot}	Mean of the T_{tot} time series	Yes	No
	Kurt_T_{tot}	Kurtosis of the T_{tot} time series	Yes	No
	Sk_T_{tot}	Skewness of the T_{tot} time series	Yes	No
	Med_T_{tot}	Median Value of the T_{tot} time series	Yes	No
	SD1_T_{tot}	Standard deviation of the points perpendicular to the line of symmetry of the Poincaré plot	Yes	No
	SD2_T_{tot}	Standard deviation of the points along the line of symmetry of the Poincaré plot	Yes	No
	SD2xSD1_T_{tot}	The multiplication of SD2_T_{tot} and SD1_T_{tot}	Yes	No
	SD2/SD1_T_{tot}	The ration between of SD2_T_{tot} and SD1_T_{tot}	Yes	No
	Rejection rate	The number of rejected T_{tot} values due to the respiration being saturated	Yes	No
$\mathbf{T}_{in}$	Median_T_{in}	Median value of the T_{in} time series	No	Yes
	SlopeInsp	Median value of the slope of the inspiration phase	Yes	No
$\mathbf{T}_{ex}$	Median_T_{ex}	Median value of the T_{ex} time series	Yes	No
Apnea	nbApnea3s	Number of apneas defined as cessation of breathing for more than three seconds	Yes	Yes
	nbApnea2cyc	Number of apneas defined as cessation of breathing with more than two missed respiratory cycles	Yes	Yes
	nbApnea3s2cyc	Number of apneas with cessation of breathing for more than three seconds and more than two missed respiratory cycles	No	Yes
	nbApnea10s	Number of apneas defined as cessation of breathing for more than ten seconds	Yes	No
Non-RRV related	GA	Gestational age	Yes	Yes

Table 4.A.1 – Category, name, and description of all RRV features included in either the LR or RFR feature set.

4.B Bradycardia Features

For the extraction of the bradycardia features, the bradycardia episodes were classified in four groups, depending on the method used for bradycardia detection and the threshold of beats per minute (BPM) considered:

— *Yellow (Yel)*: Bradycardia episodes detected by the Philips monitor and which triggered a yellow alarm.

— *Red*: Bradycardia episodes detected by the Philips monitor and which triggered a red alarm.

— *Under 100 BPM*: Bradycardia episodes detected from the ECG signal using the method described by Altuve et al. [36], and for which the heart rate fell under 100 BPM.

— *Under 80 BPM*: Bradycardia episodes detected from the ECG signal by the method previously cited, and for which the heart rate fell under a threshold 80 BPM.

All four groups of bradycardia data were segmented into six-hour periods, from which several features were extracted:

— *Length (len)*: Duration of each bradycardia episode.

— *T_min* Time between the moment when the bradycardia episode was detected, and the time of the minimum BPM for that episode.

— $T_{DiifStart}$: Time between two consecutive bradycardia episodes.

— $minBPM_{Prev}$: The difference, in BPM, between the heart rate previous to the detection of the bradycardia, and the minimum BPM during the episode.

— $minBPM_{First}$: The difference, in BPM, between the heart rate at the moment of detection of the bradycardia, and the minimum BPM during the episode.

— $Slope_{Prev}$: Describes how fast the heart rate fell from the normal value previous to the bradycardia episode, to the minimum BPM during the episode. It is calculated by the following equation:

$$\text{Slope}_{Prev} = \frac{\text{minBPMn}_{Prev}}{\text{T\_min} + 0.1\text{s}} \tag{4.4}$$

— $Slope_{First}$: Describes how fast the heart rate fell during the bradycardia episode. It is calculated by the following equation:

$$\text{Slope}_{First} = \frac{\text{minBPM}_{First}}{\text{T\_min}} \tag{4.5}$$

Finally, for each of these variables, in each of the four groups of detected bradycardia episodes, the median value and standard deviation (SD) over the six-hour segment were calculated. Those median values and SD were the features used as input for the FMA estimation algorithm described in this study.

Some types of bradycardia episodes studied might not have been present in all six-hour

segments in the database, resulting in missing or null values. Thus, some categorical features to indicate whether a value was missing or not were generated by the algorithm when handling the missing values, as explained in Section 4.2.7.

The final features sets selected by the genetic algorithm for the LR and RFR model, including both features from the original feature set, and features generated during the handling of missing data, are presented in Table 4.B.1.

Feature	**LR**	**RFR**
lenYel (median)	Yes	Yes
$T_{DiffStart}$Yel (median)	No	Yes
TminYel (median)	No	Yes
lenYel (SD)	Yes	No
minBPM$_{Prev}$Yel (SD)	No	Yes
T_minYel (SD)	Yes	Yes
lenRed (median)	Yes	Yes
T_minmRed (median)	No	Yes
lenRed (SD)	No	Yes
minBPM$_{First}$Red (SD)	Yes	Yes
T_minRed (SD)	Yes	Yes
Slope$_{Prev}$100 (median)_wasMissing	Yes	No
Slope$_{First}$100 (median)_wasMissing	No	Yes
minBPM$_{Prev}$100 (median)_wasMissing	Yes	No
$T_{DiffStart}$100 (median)_wasMissing	No	Yes
len100 (median)_wasMissing	No	Yes
minBPM$_{First}$100 (median)_wasMissing	No	Yes
Slope$_{First}$100 (SD)_wasMissing	Yes	Yes
$T_{DiffStart}$100 (SD)_wasMissing	Yes	Yes
minBPM$_{First}$100 (SD)_wasMissing	Yes	No
len100 (SD)_wasMissing	No	Yes
T_min100 (SD)_wasMissing	Yes	Yes
Slope$_{First}$80 (median)_wasMissing	Yes	Yes
Slope$_{Prev}$80 (median)_wasMissing	Yes	Yes
$T_{DiffStart}$80 (median)_wasMissing	Yes	Yes
len80 (median)_wasMissing	Yes	Yes
T_min80 (median)_wasMissing	Yes	No
minBPM$_{Prev}$80 (median)_wasMissing	No	Yes
minBPM$_{First}$80 (median)_wasMissing	No	Yes
Slope$_{First}$80 (SD)_wasMissing	Yes	Yes
Slope$_{Prev}$80 (SD)_wasMissing	No	Yes
minBPM$_{Prev}$80 (SD)_wasMissing	Yes	No
minBPM$_{First}$80 (SD)_wasMissing	Yes	No
T_min80 (SD)_wasMissing	Yes	No
GA	Yes	Yes

Table 4.B.1 – Bradycardia features included in either the LR or RFR feature set. The name of the features correspond to the name or abbreviation of the variable (len, $T_{DiffStart}$, etc.), followed by the bradycardia group (Yel, Red, 100, and 80), and in parenthesis if it is the median value or the SD. Features with *_wasMissing* at the end indicate these are categorical features added to indicate if the value in the original feature was missing or not.

Bibliography

[1] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: a review," *Medical and Biological Engineering and Computing*, vol. 44, no. 12, pp. 1031–1051, 2006.

[2] E. Longin, T. Schaible, T. Lenz, and S. König, "Short term heart rate variability in healthy neonates: normative data and physiological observations." *Early Human Development*, vol. 81, no. 8, pp. 663–671, Aug. 2005.

[3] H. Patural, V. Pichot, S. Flori, A. Giraud, P. Franco, P. Pladys, A. Beuchée, F. Roche, and J.-C. Barthelemy, "Autonomic maturation from birth to 2 years: normative values," *Heliyon*, vol. 5, no. 3, p. e01300, 2019.

[4] V. L. Schechtman, R. M. Harper, K. A. Kluge, A. J. Wilson, H. J. Hoffman, and D. P. Southall, "Heart rate variation in normal infants and victims of the sudden infant death syndrome," *Early Human Development*, vol. 19, no. 3, pp. 167 – 181, 1989.

[5] R. Statello, L. Carnevali, D. Alinovi, F. Pisani, and A. Sgoifo, "Heart rate variability in neonatal patients with seizures," *Clinical Neurophysiology*, vol. 129, no. 12, pp. 2534 – 2540, 2018.

[6] T. Al-Shargabi, R. B. Govindan, R. Dave, M. Metzler, Y. Wang, A. du Plessis, and A. N. Massaro, "Inflammatory cytokine response and reduced heart rate variability in newborns with hypoxic-ischemic encephalopathy," *Journal of Perinatology*, vol. 37, no. 6, pp. 668–672, 2017.

[7] R. Y. Moon, R. S. Horne, and F. R. Hauck, "Sudden infant death syndrome," *The Lancet*, vol. 370, no. 9598, pp. 1578 – 1587, 2007.

[8] M. H. Malloy, "Prematurity and sudden infant death syndrome: United states 2005–2007," *Journal of Perinatology*, vol. 33, no. 6, pp. 470–475, 2013.

[9] B. J. Stoll, N. Hansen, A. A. Fanaroff, L. L. Wright, W. A. Carlo, R. A. Ehrenkranz, J. A. Lemons, E. F. Donovan, J. E. T. Ann R. Stark, W. Oh., C. R. Bauer, S. B. Korones, S. Shankaran, A. R. Laptook, D. K. Stevenson, L.-A. Papile, and W. K. Poole, "Late-Onset Sepsis in Very Low Birth Weight Neonates: The Experience of the NICHD Neonatal Research Network," *Pedriatics*, vol. 110, no. 2, pp. 285–291, Aug. 2002.

[10] M.-H. Tsai, J.-F. Hsu, S.-M. Chu, R. Lien, H.-R. Huang, M.-C. Chiang, R.-H. Fu, C.-W. Lee, and Y.-C. Huang, "Incidence, clinical characteristics and risk factors for adverse outcome

in neonates with late-onset sepsis." *Pediatric Infectious Disease Journal*, vol. 33, no. 1, pp. e7–e13, Jan. 2014.

[11] G. M. Ronen, S. Penney, and W. Andrews, "The epidemiology of clinical neonatal seizures in newfoundland: A population-based study," *The Journal of Pediatrics*, vol. 134, no. 1, pp. 71 – 75, 1999.

[12] R. D. Sheth, G. R. Hobbs, and M. Mullett, "Neonatal seizures," *Journal of Perinatology*, vol. 19, no. 1, pp. 40–43, 1999.

[13] K. R. Gopagondanahalli, J. Li, M. C. Fahey, R. W. Hunt, G. Jenkin, S. L. Miller, and A. Malhotra, "Preterm hypoxic-ischemic encephalopathy," *Frontiers in Pediatrics*, vol. 4, p. 114, 2016.

[14] S. Cardoso, M. J. Silva, and H. Guimarães, "Autonomic nervous system in newborns: a review based on heart rate variability." *Childs Nervous System*, vol. 33, no. 7, pp. 1053–1063, Jul. 2017.

[15] F. A. Selig, E. R. Tonolli, E. V. C. M. d. Silva, and M. F. d. Godoy, "Heart rate variability in preterm and term neonates." *Arq Bras Cardiol*, vol. 96, no. 6, pp. 443–449, Jun. 2011.

[16] E. Longin, T. Gerstner, T. Schaible, T. Lenz, and S. König, "Maturation of the autonomic nervous system: Differences in heart rate variability in premature vs. term infants," *Journal of perinatal medicine*, vol. 34, pp. 303–8, 02 2006.

[17] H. Patural, V. Pichot, F. Jaziri, G. Teyssier, J.-M. Gaspoz, F. Roche, and J.-C. Barthelemy, "Autonomic cardiac control of very preterm newborns: a prolonged dysfunction." *Early Human Development*, vol. 84, no. 10, pp. 681–687, Oct. 2008.

[18] T. Gerstner, J. Sprenger, T. Schaible, C. Weiss, and S. Koenig, "Maturation of the autonomic nervous system: differences in heart rate variability at different gestational weeks." *Z Geburtshilfe Neonatol*, vol. 214, no. 1, pp. 11–14, Jan. 2010.

[19] T. Nakamura, H. Horio, S. Miyashita, Y. Chiba, and S. Sato, "Identification of development and autonomic nerve activity from heart rate variability in preterm infants," *Bio Systems*, vol. 79, no. 1-3, pp. 117–124, 2005.

[20] E. Helander, N. Khodor, A. Kallonen, A. Värri, H. Patural, G. Carrault, and P. Pladys, "Comparison of linear and non-linear heart rate variability indices between preterm infants at their theoretical term age and full term newborns," in *EMBEC & NBC 2017*, H. Eskola, O. Väisänen, J. Viik, and J. Hyttinen, Eds. Singapore: Springer Singapore, 2018, pp. 153–156.

[21] K. L. Fyfe, S. R. Yiallourou, F. Y. Wong, A. Odoi, A. M. Walker, and R. S. C. Horne, "The effect of gestational age at birth on post-term maturation of heart rate variability." *Sleep*, vol. 38, no. 10, pp. 1635–1644, Oct. 2015.

[22] S. B. Mulkey, S. Kota, C. B. Swisher, L. Hitchings, M. Metzler, Y. Wang, G. L. Maxwell, R. Baker, A. J. du Plessis, and R. Govindan, "Autonomic nervous system depression at term in neurologically normal premature infants." *Early Human Development*, vol. 123, pp. 11–16, Aug. 2018.

[23] C. D. Smyser, N. U. Dosenbach, T. A. Smyser, A. Z. Snyder, C. E. Rogers, T. E. Inder, B. L. Schlaggar, and J. J. Neil, "Prediction of brain maturity in infants using machine-learning algorithms," *NeuroImage*, vol. 136, pp. 1 – 9, 2016.

[24] N. J. Stevenson, L. Oberdorfer, N. Koolen, J. M. O'Toole, T. Werther, K. Klebermass-Schrehof, and S. Vanhatalo, "Functional maturation in preterm infants measured by serial recording of cortical activity," *Scientific Reports*, vol. 7, no. 1, p. 12969, 2017.

[25] M. Chiera, F. Cerritelli, A. Casini, N. Barsotti, D. Boschiero, F. Cavigioli, C. G. Corti, and A. Manzotti, "Heart rate variability in the perinatal period: A critical and conceptual review." *Frontiers in neuroscience*, vol. 14, p. 561186, 2020.

[26] F. Rusconi, M. Castagneto, L. Gagliardi, G. Leo, A. Pellegatta, N. Porta, S. Razon, and M. Braga, "Reference values for respiratory rate in the first 3 years of life." *Pediatrics*, vol. 94, no. 3, pp. 350–355, Sep. 1994.

[27] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. Maconochie, L. Tarassenko, and D. Mant, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies." *The Lancet*, vol. 377, no. 9770, pp. 1011–1018, Mar. 2011.

[28] S. A. Lorch, L. Srinivasan, and G. J. Escobar, "Epidemiology of apnea and bradycardia resolution in premature infants," *Pediatrics*, vol. 128, no. 2, pp. e366–e373, 2011.

[29] M. A. Mohr, K. D. Fairchild, M. Patel, R. A. Sinkin, M. T. Clark, J. R. Moorman, D. E. Lake, J. Kattwinkel, and J. B. Delos, "Quantification of periodic breathing in premature infants," *Physiological measurement*, vol. 36, no. 7, pp. 1415–1427, 07 2015.

[30] B. R. Greene, P. de Chazal, G. Boylan, R. B. Reilly, C. O'Brien, and S. Connolly, "Heart and respiration rate changes in the neonate during electroencephalographic seizure," *Medical and Biological Engineering and Computing*, vol. 44, no. 1, pp. 27–34, 2006.

[31] M. S. Miller, K. M. Shannon, and G. T. Wetzel, "Neonatal bradycardia," *Progress in Pediatric Cardiology*, vol. 11, no. 1, pp. 19 – 24, 2000.

[32] "Age terminology during the perinatal period," *Pediatrics*, vol. 114, no. 5, pp. 1362–1364, 2004.

[33] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *Journal of Econometrics*, vol. 187, no. 1, pp. 95–112, Jul. 2015.

[34] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Computer Surveys*, vol. 45, no. 1, Dec. 2012.

[35] X. Navarro, F. Porée, A. Beuchée, and G. Carrault, "Artifact rejection and cycle detection in immature breathing: Application to the early detection of neonatal sepsis," *Biomedical Signal Processing and Control*, vol. 16, pp. 9 – 16, 2015.

[36] M. Altuve, G. Carrault, A. Beuchée, P. Pladys, and A. I. Hernández, "On-line apnea-bradycardia detection using hidden semi-markov models," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 4374–4377.

[37] M. Doyen, D. Ge, A. Beuchée, G. Carrault, and A. I. Hernández, "Robust, real-time generic detector based on a multi-feature probabilistic method," *PLOS ONE*, vol. 14, no. 10, pp. 1–22, Oct. 2019.

[38] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Frontiers in Public Health*, vol. 5, p. 258, 2017.

[39] M. Bolanos, H. Nazeran, and E. Haltiwanger, "Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2006, pp. 4289–4294.

[40] A. Bauer, J. W. Kantelhardt, P. Barthel, R. Schneider, T. Mäkikallio, K. Ulm, K. Hnatkova, A. Schömig, H. Huikuri, A. Bunde, M. Malik, and G. Schmidt, "Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study," *The Lancet*, vol. 367, no. 9523, pp. 1674–1681, May 2006.

[41] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: A survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007.

[42] M. E. J. Newman, "Assortative Mixing in Networks," *Physical Review Letters*, vol. 89, no. 20, Oct. 2002.

[43] M. Zhao, C. Fu, L. Ji, K. Tang, and M. Zhou, "Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5197–5204, 2011.

[44] R. H. H. Groenwold, "Informative missingness in electronic health record systems: the curse of knowing," *Diagnostic and Prognostic Research*, vol. 4, no. 1, p. 8, 2020.

[45] X. Sun, Y. Zhao, and X. Xue, "Analyzing spatial characters of the ecg signal via complex network method," in *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, vol. 3, Oct. 2011, pp. 1650–1653.

[46] T. Nguyen Phuc Thu, A. I. Hernández, N. Costet, H. Patural, V. Pichot, G. Carrault, and A. Beuchée, "Improving methodology in heart rate variability analysis for the premature infants: Impact of the time length," *PLOS ONE*, vol. 14, no. 8, pp. 1–14, 08 2019.

CHAPTER 5

Evaluation of the Ensemble Machine Learning Model on a Population of Preterm Infants with Abnormal Maturation

5.1 Introduction

In Chapter 4 we proposed an automated method for the evaluation of maturation in healthy premature infants based on an ensemble machine learning (EML) model. The proposed model works by taking as input physiological signals and estimating a functional maturational age (FMA) based on that information. We showed that this FMA had a small error in comparison to the postmenstrual age (PMA) of the infants, and hypothesized that this error should be more significant in a population of infants with abnormal maturation.

We base this hypothesis on the findings from previous studies, which were discussed at length in Section 1.2.2. Some of these studies were able to determine, based on magnetic resonance imaging of the brain from the perinatal period, which infants would present cognitive [1], motor [2], or language [3] impairment at two years of age. This suggests that information acquired during the perinatal period and neonatal intensive care unit (NICU) hospitalization can shed light on the long-term maturation and development of preterm infants.

Another series of studies, proposed by Stevenson et al., also discussed in more detail in Section 1.2.2, had an approach more similar to ours, by also estimating a functional maturational age, but based on electroencephalogram (EEG) data, and comparing it to the PMA. In the first of these studies [4] they developed a machine learning model to estimate the PMA based on the EEG data from healthy preterm babies who had normal outcomes at 12 months of age. In their next study [5], Stevenson et al. included premature infants with abnormal neurodevelopmental outcome at 12 months of age, and showed that they presented a larger difference between the maturational age estimated from the EEG data acquired during the perinatal period and the PMA than those with normal outcomes.

The findings from these studies further suggest that abnormal maturational outcomes in preterm infants can be detected in the postnatal period, and that comparing a functional maturational age estimated through machine learning models from physiological data, to the clinically determined PMA of the infants, might serve as an objective metric to detect maturational delays. This is consistent with the results we obtained on Chapter 4, where we showed we were able

to estimate a FMA based on heart rate variability (HRV), respiration rate variability (RRV), and bradycardia data, with a mean absolute error (MAE) that ranged from 0.93 to 1.39 weeks compared to the PMA.

Consequently, in this chapter we aim to test our hypothesis that the difference between the FMA estimated from physiological data and the PMA will be bigger for a population of infants who had abnormal maturation during their stay in NICU. To do this we will train the model designed in Chapter 4 in a population of premature infants who did not present any complications or signs of abnormal maturation during their NICU stay. Then, we will test the model on both healthy preterm babies with no obvious signs of delayed maturation during their hospitalization, and on a population of infants who presented complications associated with delayed maturation, to then compare the results obtained for each population. Similarly to the previous chapter, we use HRV, RRV, and bradycardia features.

In the next sections we describe the population of infants used in this study, and how it was classified into healthy and unhealthy populations, based on a number of conditions associated with negative neurodevelopmental outcomes. We then explain how the data was split in training and testing set. Afterwards we explain the evaluation metrics used to describe the results, and then detail the results obtained by applying the model described in Chapter 4 on this train and test set and compare the results obtained on the healthy test patients, to the ones obtained for the unhealthy population. Finally, we discuss these results and their significance.

5.2 Constitution and Classification of the Study Population

The data used in this study is also part of the database of the Digi-NewB cohort (NCT02863978, EU GA n°689260), which prospectively included infants born between 25 and 42 weeks of gestation, hospitalized in the NICU of six university hospitals in western France (University Hospitals of Rennes, Angers, Nantes, Brest, Poitiers, and Tours) in 2017-2019. The collection of data was carried out after approval by the ethics committee (CPP Ouest 6-598) and informed parental consent.

The population of healthy infants used in this study is exactly the same as described in Section 4.2.1, for all three datasets: HRV, RRV, and Bradycardia. Therefore the selection criteria was the same as described in said section, and the data acquisition, processing and feature extraction are also the same as described in Chapter 4. Although this population includes infants born prematurely, which by its very nature is associated to inherent medical complications and developmental delays compared to infants born at term, we will refer to this population as the healthy population throughout this chapter, to differentiate it from the population of infants who presented additional complications and who presented a delayed maturation even compared to other infants born at a similar GA.

Data Type		HRV Features					
Train/Test Set		Train	Test				
Term	GA (weeks)	Healthy (n = 40)	Healthy (n= 10)	Abnormal Maturation (n = 54)			
				NL (n = 6)	BPD (n = 24)	NEC (n = 2)	Multi. (n = 22)
EP	[24, 28[	7	2	2	12	1	13
VP	[28, 32[	11	3	4	12	1	8
LP	[32, 37[	10	2	0	0	0	1
ET	[37, 39[	5	1	0	0	0	0
FT	≥39	7	2	0	0	0	0

Table 5.2.1 – Distribution by age and classification of the population in the HRV dataset. The *Multi.* column refers to the infants who presented multiple conditions from the inclusion criteria for the unhealthy population.

For the population of infants who presented abnormal maturation, the selection was done by the same medical experts as for the healthy infants. The inclusion criteria were premature infants who were diagnosed with at least one of the following conditions during their NICU hospitalization: (i) interventricular hemorrhage (IVH) [6]; (ii) periventricular leukomalacia (PVL) [7]; (iii) bronchopulmonary dysplasia (BPD) [8]; (iv) necrotizing enterocolitis (NEC) [9]. These conditions where chosen as the inclusion criteria because all of them have been associated with negative long-term neurodevelopmental outcomes for infants ([10, 11, 12, 13, 14, 15, 16]).

5.2.1 Classification of the Study Population Based on Data Type

For this study we will use three datasets: HRV features dataset, RRV features dataset, and bradycardia dataset. The data acquisition, data processing, and feature extraction for each dataset was done in the same manner as described in Chapter 4. Furthermore, as the feature selection was already done in the previous study, here we will only use those features that the genetic algorithm chose as optimal for each of the types of the dataset. These features were shown in Tables 4.3.1, 4.A.1, 4.B.1, respectively.

It is important to note that that all three datasets were built using the same population of infants, both for the healthy population and the population with abnormal maturation. Therefore, any differences in the number of infants in each of these datasets is due to the exclusion of neonates which did not have enough recordings of good quality in the case of the RRV population; or, in the case of the bradycardia population, because their data had not been processed yet for the detection of bradycardia episodes and the extraction of the associated features.

The information regarding number of infants in each dataset, as well as how these infants are distributed in terms of their gestational age (GA) and medical history, is presented in Table 5.2.1 for the HRV dataset, Table 5.2.2 for the RRV dataset, and 5.2.3 for the bradycardia features dataset.

Data Type		**RRV Features**					
Train/Test Set		**Train**	**Test**				
Term	**GA (weeks)**	**Healthy (n = 38)**	**Healthy (n = 10)**	**Abnormal Maturation (n = 54)**			
				NL (n = 6)	**BPD (n = 24)**	**NEC (n = 2)**	**Multi. (n = 22)**
EP	[24, 28[	7	2	2	12	1	13
VP	[28, 32[	11	3	4	12	1	8
LP	[32, 37[	10	2	0	0	0	1
ET	[37, 39[	3	1	0	0	0	0
FT	≥39	7	2	0	0	0	0

Table 5.2.2 – Distribution by age and classification of the population in the RRV dataset. The *Multi.* column refers to the infants who presented multiple conditions from the inclusion criteria for the unhealthy population.

Data Type		**Bradycardia Features**					
Train/Test Set		**Train**	**Test**				
Term	**GA (weeks)**	**Healthy (n = 35)**	**Healthy (n = 8)**	**Abnormal Maturation (n = 52)**			
				NL (n = 6)	**BPD (n = 24)**	**NEC (n = 2)**	**Multi. (n = 20)**
EP	[24, 28[	7	2	2	12	1	12
VP	[28, 32[	11	3	4	12	1	7
LP	[32, 37[	9	2	0	0	0	1
ET	[37, 39[	3	0	0	0	0	0
FT	≥39	5	1	0	0	0	0

Table 5.2.3 – Distribution by age and classification of the population in the Bradycardia dataset. The *Multi.* column refers to the infants who presented multiple conditions from the inclusion criteria for the unhealthy population.

5.2.2 Classification of the Study Population Based on Gestational Age

The population was split into five categories, according to the gestational GA of the infant: extreme preterm (EP), very preterm (VP), late preterm (LP), early term (ET), and full term (FT). The cut-off GA for each group was determined based on the recommendations for preterm birth [17] and for term birth [18] classification. The same rule was applied to every data type (HRV, RRV, and bradycardia) and to both the healthy and unhealthy populations. This is also consistent with how the classification by term and GA was done in Chapter 4.

The cut-off GA for every term group as well as the number n of infants in each group, for both healthy and unhealthy population, in the HRV, RRV, and Bradycardia datasets is detailed in Tables 5.2.1, 5.2.2, and 5.2.3, respectively.

The definition of GA, PMA, and chronological age that we use throughout this article is that proposed by the American Academy of Pediatrics [19].

5.2.3 Classification of the Study Population Based on Medical History

For the purpose of this study, the infants included in the population with abnormal maturation (PAM) were further split into four categories. The first category is infants who presented only IVH, PVL, or both, which we refer to as the neurological lesion (NL) population. The second category is infants who only presented BPD, which we refer to as the BPD population. The third category contains the infants who only presented NEC, which we refer to as the NEC population. Finally, in the fourth category are all the infants who presented more than one of the conditions listed in the inclusion criteria. We did this classification in order to be able to evaluate if the model is more sensitive to some conditions than to others.

The fifth category under the classification based on medical history are the infants belonging to the healthy population, who were considered to have a normal maturation pattern for infants born at their GA. These, as we mentioned before, are the same infants included in the population used for the study presented in Chapter 4, and the inclusion criteria for this population was presented in Section 4.2.1.

5.2.4 Classification of the Study Population into Train Set and Test Set for EML Model

The main difference in the methodology we will use in this chapter as compared to the previous one is regarding how the EML model is trained and tested. In the study presented in Chapter 4, we trained and tested the proposed model using the leave-one-out cross-validation technique. However, although we retain the features that were selected as explained in Chapter 4, and we will use the same ensemble machine learning model architecture, in the present study we will use a train/test split of the data and retrain the model on the training data.

We chose to use the train/test method in this occasion because we want to have one single trained model for each data type, on which we can then test both the healthy test population and the population of neonates with abnormal maturation. Using the leave-one-out cross-validation method would instead yield a slightly different trained model for every healthy patient in the population, making harder the comparison between results from the healthy and unhealthy population.

Therefore, for the train set we have retained 80% of the healthy population in each dataset. These were chosen randomly, but choosing the same percentage from every term group to avoid bias towards any GA. Given that our hypothesis is that the model will estimate FMAs that differ more from the clinically determined PMA for the infants that presented abnormal maturation in comparison to those that were healthy during their NICU stay, no infants from the population with abnormal maturation were used to train the model, as this could introduce a bias in the model.

For the test set we retained the remaining 20% of the healthy population in each dataset. Also, all the infants in the abnormal maturation populations were used exclusively as test patients.

5.3 Evaluation Metrics

In order the evaluate and compare the results obtained on the healthy test population to those obtained in the population with complications related to abnormal neurodevelopmental outcomes, we will calculate the mean, maximum, and minimum MAE for each category of the population in the test set, as well as its standard deviation (SD) and 95% confidence interval (CI). These are the same metrics we used in the previous chapter for the evaluation of the model's performance.

Additionally, to be able to objectively compare the results for each category of the population in the test set we used the repeated measures correlation described in [20], and its associated R language package [21]. We used this to study the correlation between the PMA and the FMA estimated by the model. This type of correlation is specially well suited for cases such as this, where each population might have a different number of patients, and the number of measurements for each patient may vary, but every measurement is paired. This is so because for every measurement of the FMA we also have its associated PMA, and vice versa.

The result of the repeated measures correlation is interpreted similarly to other correlation metrics: a correlation of -1 indicates that the data has a perfect negative intra-individual association between the variables; a correlation of 0 signals that the variables has no intra-individual association; a correlation equal to 1 indicates that the data has a perfect positive intra-individual association between the variables. Each correlation also has an associated p-value.

5.4 Results

In this section we present the results obtained by using the EML model we proposed in Chapter 4 on the test set of each dataset of the population described in Section 5.2. Specifically, we detail the results for the healthy test patients and for all the patients in the abnormal maturation population, both for the entire population with abnormal maturation (PAM), and for each of the categories into which we classified this population, as detailed in section 5.2.3.

5.4.1 Performance of the Model on Healthy and Unhealthy Population using HRV Data

In Table 5.4.1 we present the results obtained in the HRV test set. We observe that the healthy population has a very low MAE, of 0.92 weeks, and has a very strong correlation between the PMA and FMA (0.91, p-value < 0.0001). All the other categories in the test population have

	Mean MAE (weeks)	Max. MAE (weeks)	Min. MAE (weeks)	SD (weeks)	95% CI (weeks)	Correlation	p-value
Healthy (n = 10)	0.92	1.53	0.04	0.48	[0.57, 1.26]	0.91	< 0.0001
NL (n = 6)	1.91	3.31	0.56	1.07	[0.78, 3.03]	0.459	< 0.0001
BPD (n = 24)	1.14	2.33	0.05	0.78	[0.81, 1.48]	0.832	< 0.0001
NEC (n = 2)	2.50	2.65	2.36	0.2	[2.36, 2.65]	0.685	0.0009
Multi. (n = 22)	2.09	4.83	0.63	1.12	[1.59, 2.58]	0.705	< 0.0001
PAM (n = 54)	1.66	4.83	0.05	1.05	[1.38, 1.95]	0.74	< 0.0001

Table 5.4.1 – Performance of the model on the HRV test set, as measured by the mean, maximum, and minimum MAE, and its SD, 95% CI, and correlation between PMA and FMA and its p-value.

a lower correlation and higher mean MAE, maximum MAE (Max. MAE), and minimum MAE (Min. MAE). The SD and the range of the 95% CI is also smaller for the healthy population than for any other, except for the NEC population. This fact is probably due to the NEC population having only two infants.

The bigger dispersion in the FMA values for the unhealthy population is also noticeable in the scatter plot presented in Figure 5.4.1, where it can be observed that the healthy population (in orange) tends to have results that are closer to the line for which the PMA and FMA are the same, with only a few outliers. In comparison, all the populations with abnormal maturation show more dispersion, displaying mostly an overestimation of the FMA in comparison with the PMA between weeks 24 and 28. This is followed by an underestimation of the FMA in comparison with the PMA, which becomes more accentuated starting on the 32nd week.

In the table we can also observe that the population with the least correlation between PMA and FMA is the NL population, with a correlation of 0.46 (p-value < 0.0001). This population also has a high mean MAE, at 1.91 weeks. This is almost one week more than the healthy population. From all the subcategories of the unhealthy population, the one with the highest correlation between PMA and FMA is the BPD population, with 0.83. This population also has the lowest mean MAE among the population with abnormal maturation.

5.4.2 Performance of the Model on Healthy and Unhealthy Population using RRV Data

	Mean MAE (weeks)	Max. MAE (weeks)	Min. MAE (weeks)	SD (weeks)	95% CI (weeks)	Correlation	p-value
Healthy (n = 10)	1.44	4.12	0.106	1.26	[0.54, 2.34]	0.655	< 0.0001
NL (n = 6)	2.06	2.91	1.24	0.54	[1.49, 2.62]	0.153	0.046
BPD (n = 24)	1.63	3.36	0.18	0.82	[1.28, 1.97]	0.461	< 0.0001
NEC (n = 2)	3.05	4.15	1.96	1.55	[1.96, 4.15]	0.453	0.068
Multi. (n = 22)	2.68	5.47	0.90	1.24	[2.13, 3.23]	0.333	< 0.0001
PAM (n = 54)	2.16	5.47	0.18	1.11	[1.85 2.46]	0.37	< 0.0001

Table 5.4.2 – Performance of the model on the RRV test set, as measured by the mean, maximum, and minimum MAE, its SD, 95% CI, and correlation between PMA and FMA and its p-value.

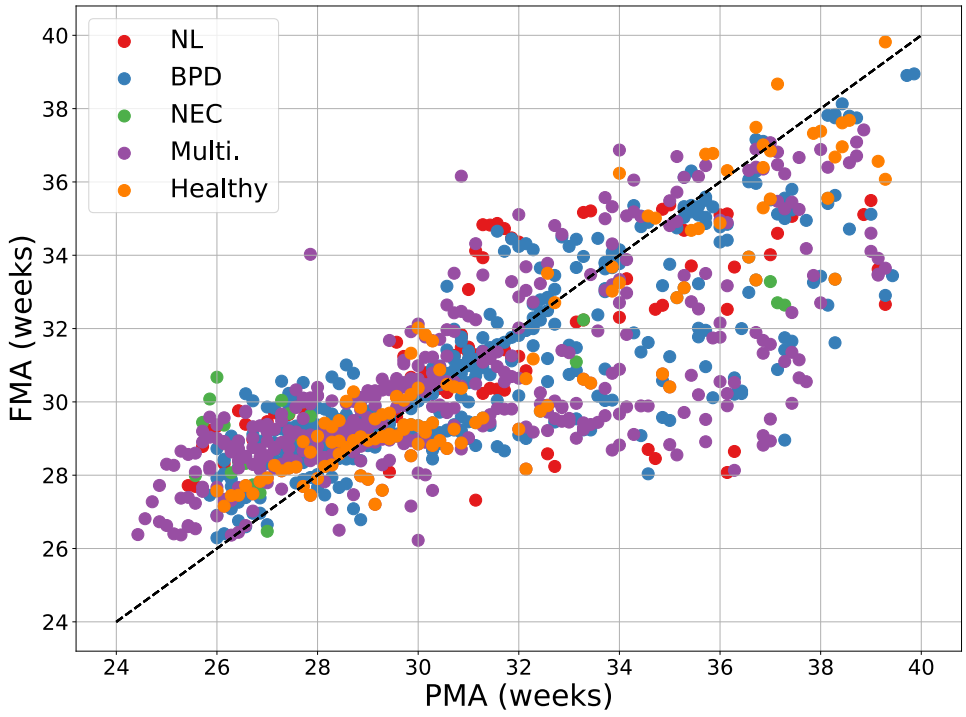


Figure 5.4.1 – PMA versus FMA for all infants in the HRV test set, grouped by population classification based on medical history

Table 5.4.2 shows the results for the RRV test set. We observe that once again the healthy population has the highest correlation between PMA and FMA and the lowest mean and minimum MAE. From the unhealthy infants, those belonging to the BDP subcategory are the ones with the highest correlation between PMA and FMA (0.461), and lowest mean and minimum MAE. On the opposite end of the spectrum is the NL population, with a very low correlation between PMA and FMA (0.153, p-value = 0.046), followed by infants that presented multiple conditions (Multi.), for whom the correlation is 0.333 (p-value < 0.0001).

In Figure 5.4.2 we observe that the healthy infants (in orange) seem to be generally closer to the diagonal for which FMA and PMA would be equal, while the population with multiple conditions (in purple) seems to deviate from this line.

5.4.3 Performance of the Model on Healthy and Unhealthy Population using Bradycardia Data

The results obtained on the Bradycardia test population are presented in Table 5.4.3. In this case we also observe that the healthy population presents the smallest mean MAE (1.53), and the biggest correlation between PMA and FMA (0.602, p-value < 0.0001), followed closely by

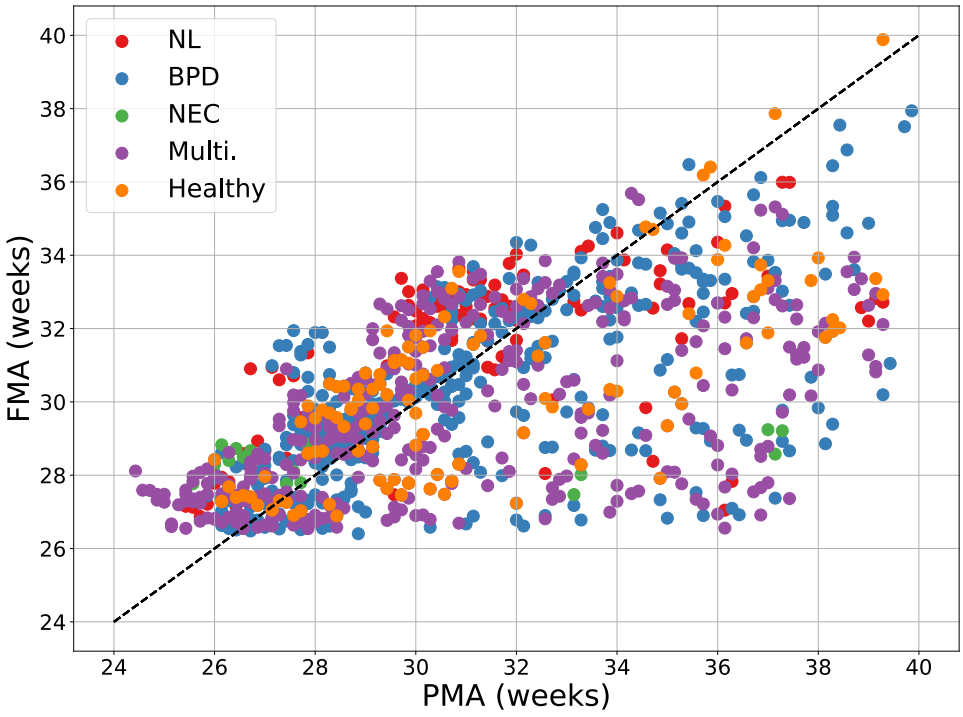


Figure 5.4.2 – PMA versus FMA for all infants in the RRV test set, grouped by population classification based on medical history

	Mean MAE (weeks)	Max. MAE (weeks)	Min. MAE (weeks)	SD (weeks)	95% CI (weeks)	Correlation	p-value
Healthy (n = 10)	1.53	2.76	0.02	0.88	[0.80, 2.26]	0.638	< 0.0001
NL (n = 6)	1.77	2.31	1.31	0.44	[1.31, 2.23]	0.353	0.0015
BPD (n = 24)	1.85	9.09	0.16	1.81	[1.10, 2.60]	0.602	< 0.0001
NEC (n = 2)	2.98	3.51	2.45	0.88	[2.45, 3.51]	0.178	0.525
Multi. (n = 20)	2.76	6.37	0.5	1.57	[2.03, 3.49]	0.341	< 0.0001
PAM (n = 52)	2.31	8.73	0.29	1.55	[1.89, 2.74]	0.441	< 0.0001

Table 5.4.3 – Performance of the model on the Bradycardia test set, as measured by the mean, maximum, and minimum MAE, its SD, 95% CI, and correlation between PMA and FMA and its p-value.

the BDP population with correlation 0.602 (p-value < 0.0001).

The NEC population is the one with the highest mean MAE, at 2.98 weeks, and lowest correlation, at 0.178, however the p-value associated with the correlation is very high, at 0.525, suggesting that this result is not statistically significant. This is not surprising, given the small size of this population (n = 2). The population with the next highest mean MAE and lowest correlation, is the population with multiple conditions (Multi.), with correlation 0.341 (p-value < 0.0001) and mean MAE of 2.76 weeks.

In Figure 5.4.3 we present the scatter plot of PMA versus FMA for this population. We

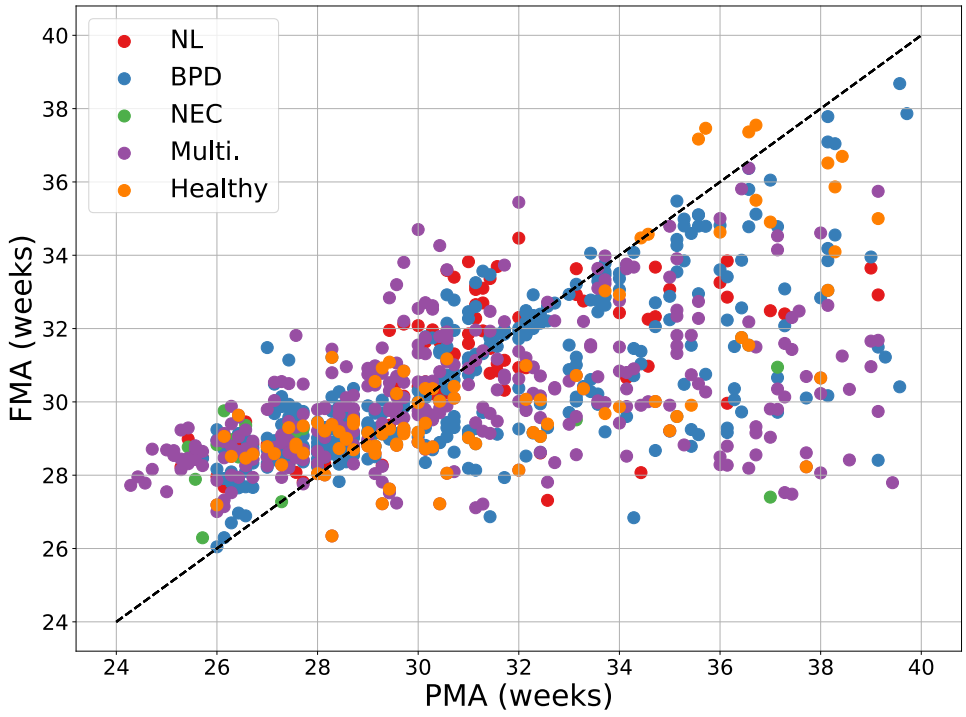


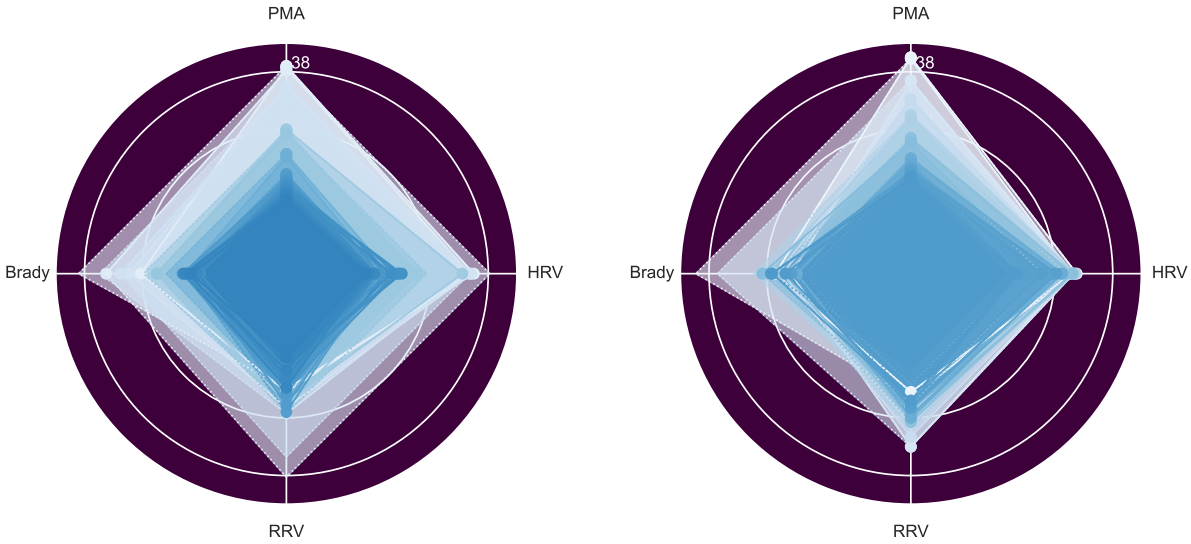
Figure 5.4.3 – PMA versus FMA for all infants in the Bradycardia test set, grouped by population classification based on medical history

observe that the healthy population (in orange), even though it presents some outliers, tends to have its results closer to the diagonal for PMA equal to the FMA, while the subcategories of the PAM present more dispersion.

5.5 Sample Cases

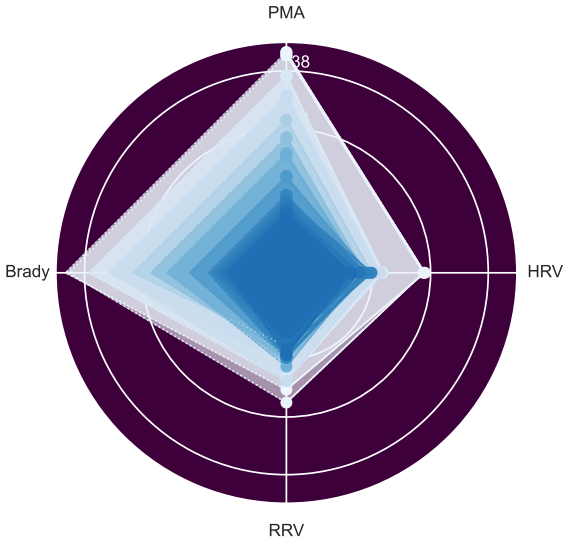
In Figure 5.5.1 the results from four infants from the test set, to exemplify how the results could be presented and analysed on an individual basis. The type of spider graph presented in the figure, which was designed and implemented by other members of the Digi-NewB team, is an alternative to the scatter plot visualization we presented for sample cases in Chapter 4. This graph allows the presentation of the results along different axes, each one of them displaying the functional maturational age related to a specific data type. Another axis displays the PMA, in order to facilitate the interpretation of the results. These graphs could be easily adapted to include more axis, if more data types were analysed, and they can also handle cases there are data points missing for any of the axis.

In the graphs presented Figure 5.5.1, there are three axis corresponding to FMAs, one axes for the FMA based on each data type we studied, namely HRV, RRV, and bradycardia (presented

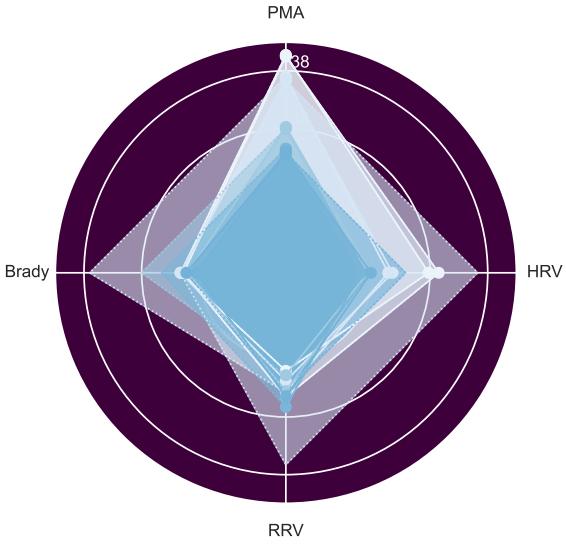


(a) VP infant from the healthy population.

(b) VP infant from the NL population.



(c) EP infant with BPD and NEC.



(d) EP infant with NL, BPD, and NEC.

Figure 5.5.1 – Estimated FMA along each axes for sample cases from the test set.

as *Brady* in the graph). The fourth axes represents the PMA. The larger white circumference corresponds to 38 weeks for any axes, while the smaller white circle corresponds to 34 weeks. For any given PMA for which a FMA was estimated in at least one of the FMA axis, a quadrilateral is drawn, with its vertexes anchored in each of the axis. In this figure we present the results from multiple days in the same graph. If the FMA for a certain axis was calculated for that day (corresponding to a specific PMA), the vertex in the corresponding axes is marked with a circle, in the place that corresponds to the value of the estimated FMA, and solid lines are used to

draw the sides of the quadrilateral connecting it to other vertexes. If the FMA for a certain axis corresponding that day is missing, then for visualization purposes, the vertex is marked without the circle, and is placed in the value along the axes that corresponds to the true PMA for that day, with dotted lines connecting it to the neighboring vertex.

Figure 5.5.1a shows the results obtained for an infant in the healthy population of the test set. This is a very preterm female infant, born at 29 weeks of gestational age, and who was considered by the neonatologists in the team to have a normal maturational trajectory for infants born at a similar GA. We observe that for this patient, the quadrilaterals in the graph corresponding to different days all have approximately square shapes, which suggests that the FMAs calculated from each data type were similar, and that they were also similar to the PMA. We observe that the last data point corresponds to approximately 38 weeks of PMA, the FMAs corresponding to this day along the HRV and bradycardia (denoted as *Brady)* axis, even though they are bellow 38 weeks, they are still reasonably close to the 38 weeks, which seems to suggest there is no major neurodevelopmental delay in this infant. Along the RRV axis we observe that, while there were several days with data available, the data corresponding to the last two days with measurements did not have RRV data, this is denoted by the type of vertex drawn and the dashed lines.

In Figure 5.5.1b we show the results for a VP female infant born at 30 weeks and 5 days of GA. This infant belonged to the NL population, who was diagnosed with IVH. For this infant we observe that the shape of the quadrilateral is more irregular, with the highest PMA value, which is slightly over 38 weeks, corresponding to very low values along all FMA axis, where the value is closer to the circle that corresponds to 34 weeks. This suggest that the FMA is considerably lower than the PMA for all the data types we analysed.

Figure 5.5.1c corresponds to the results from an EP female patient born at 28 weeks and 6 days of GA. This patient belonged to the population with multiple conditions. Specifically, she was diagnosed with BPD and NEC. We observe that this is one of the patients for whom there was no bradycardia available. In the graph, this is denoted by the drawing of the vertex itself, which is not marked with a circle, and the dashed lines that connect it to the neighboring vertexes. Therefore, for the bradycardia axis, each vertex is placed in the value corresponding to the PMA. On the other hand, for the HRV and RRV, there are a lot of data points, which allows to observe how for the earlier days of life (represented by the darker blue quadrilaterals closer to the center), the PMA and the FMAs estimated from the HRV and RRV data were very similar, resulting in the corresponding quadrilaterals being nearly square. However, in later days, as the PMA continues to increase, the HRV and RRV FMAs lagged behind, and never measured more than 34 weeks, even when the last measurements correspond to a PMA of more than 38 weeks.

Finally, in Figure 5.5.1d we present the results from a VP male infant, born at 28 weeks and 3 days of GA. This infant was also in the population with multiple conditions, and was in fact diagnosed with NL, BPD. and NEC. For this infant we observe that the quadrilaterals

drawn for each day have very irregular shape. Also, the estimated FMAs have a very erratic behaviour, as some of the FMAs corresponding to a higher PMA have lower values than the FMAs corresponding to a lower PMA. This is particularly clear in the RRV axis, where we can observe that while the FMA corresponding to approximately 34 weeks of PMA was estimated with a very small error, the FMA estimated for a PMA of over 38 weeks was considerably below 34 weeks and also below the FMA estimated at 34 weeks PMA. However, and even though the HRV FMA was also severely underestimated, with at least four weeks of error, the biggest underestimation was for the bradycardia FMA. For this data type, the estimated FMA never reached the 34 weeks, even when the last PMA was over 38 weeks.

5.6 Discussion

In this chapter we applied the ensemble machine learning model we developed in Chapter 4 to try and estimate the functional maturational age of neonatess who presented complications during their NICU hospitalization, which have been documented to have a negative impact in the short and long-term neurodevelopmental outcomes of the infants. We referred to this population as the population with abnormal maturation, or PAM, and our hypothsesis was that the FMA estimated for these infants would differ more from their PMA than for a population of healthy infants. To test this hypothesis, we retrained the EML model on a population of only healthy infants who did not present any health complications or signs of abnormal maturation for infants of born at their gestational age. We estimated the FMA for a test population of healthy infants and for the PAM, and calculated the difference between PMA and FMA in terms of the MAE and the correlation between the two. To determine if the model was more sensitive to some medical conditions than other, we also did this analysis for subcategories of the PAM.

Our main finding was that, as hypothesized, the healthy test population displayed a lower MAE and a higher correlation between PMA and FMA than the PAM and than any subcategory of the PAM. This was true for each of the datasets we tested, which were the HRV, RRV, and Bradycardia datasets. The results obtained in this population were also consistent with the results we obtained in Chapter 4, where we only focused on a healthy population for the purpose of developing and validating the model.

We also observed that, of the three datasets, HRV had the least MAE and the highest correlation for the healthy infants, with the Bradycardia dataset being in the opposite end of the spectrum. This might be due to a combination of factors. The first being that there are differences in the size of the train set for each dataset due to the available data. The HRV train set was the biggest (n = 40), followed by the RRV train set (n = 38), and bradycardia being the smallest (n = 35). This might cause the model to perform better on HRV data just by having more examples to train on and thus being able to generalize better. The other contributing

factor might be that, as discussed in Chapter 4, RRV and Bradycardia features are less strongly correlated to the PMA than HRV features. Therefore, these data types might need even larger datasets for the model to perform as well as it does for HRV.

Regardless of these limitations, the model still seems to be sensible to disruptions in the maturation of the infants for all data types, as suggested by the fact that despite differences in performance between each data type, for all of them the healthy population had an estimated FMA closer to the PMA than any other test population.

However, the behaviour of the PAM varied slightly between datasets. For both the HRV and the RRV datasets, the subcategory of the PAM that showed the most disruption, both in terms of the lowest correlation between FMA and PMA was the population with neurological lesions (NL), which included infants who presented IVH or PVL. In the HRV dataset this was followed by the NEC population, while in the RRV population, the population with next lower correlation were the infants that displayed multiple conditions. For the Bradycaria set, however, the largest disruption, in terms of the lowest correlation and the highest MAE was the NEC population, although the results in this population might be particularly susceptible to noise in the data given that it only has two patients. The population with the next most disruption were the infants with multiple conditions, followed by the NL population, with correlations of 0.341 and 0.353 respectively. However, for all three datasets the condition that seemed to cause the less disruption was BPD.

Finally, by presenting the results from four sample cases we were able to show one possible way in which the FMA from different data types and their corresponding PMA, could be displayed graphically for individual patients. This allowed us to demonstrate how the model could be implemented as DSS and how the estimated FMA along different axis could be interpreted.

These results suggest that the proposed method can be applied to preterm infants in NICU settings to estimate their functional maturational age based on physiological signals and their GA. With a bigger difference between an infant's PMA and the FMA estimated by the EML model being interpretable as a disruption in the maturation pattern of that infant. This could be implemented as a decision support system, helping physicians, for instance, to better assess when to stop supported ventilation, when to discharge the patient, and implement early intervention and support for the patients and their families to handle potential long-term neurodevelopmental negative outcomes and possible help minimize their impact.

5.7 Conclusion

In this chapter we confirmed our hypothesis that the method developed and presented in Chapter 4 could potentially detect disruptions in the maturation of infants in the postnatal period. The method employed uses automated feature selection, by applying filtering and a

genetic algorithm, and an ensemble machine learning model that combines linear regression and random forest regression. It was designed and implemented in such a way that models can be easily trained and tested on different data types.

For the study presented in this chapter we trained the model with a population of healthy infants, and then tested it on a test set that included that healthy infants as well as infants that were diagnosed during the postnatal period with medical conditions associated with negative neurodevelopmental outcomes. We found that the estimated FMA was closer to the PMA for the healthy infants, both in terms of a lower mean absolute error, and a higher correlation, than it was for the infants with abnormal maturation. This result was consistent across all the data types we studied, which were HRV, RRV, and bradycardia. The results were not as uniform for the population with abnormal maturation and its subcategories. For the HRV and RRV datasets we found that the most disruptive condition seems to be neurological lesions, which included IVH and PVL, while for the bradycardia dataset the most disruptive were necrotizing enterocolitis, followed by the combination of multiple conditions. While the least disruptive condition, for all data types included in the study, seemed to be bronchopulmonary dysplasia.

Additionally, it is worth mentioning as an important result that we also propose an interface that describes the maturation of the premature infants along the three axis we evaluated. The interest of this representation is to give a quick and visual feedback to the clinician about the maturation of the infants during their stay in the NICU.

The results presented here suggest that machine learning models trained on physiological data which is continuously and routinely available in NICU settings can be used can evaluate the maturational progress of the infants in a non-invasive manner, in terms of calculating a functional maturational age. While we hope that the model presented in this dissertation can serve as proof of concept to motivate further research, the model needs to go through further validation. This should include testing the model on more data types available in the project, such as motion, cry, and sleep data. Finally, the model should also be validated with more data for both the healthy and unhealthy populations in order to build more reliable and robust models, with the final goal of producing an integrated decision support system that can, in real-time, help physicians evaluate the maturation process of the infants in different axes based on all the data available.

Bibliography

[1] K. Schadl, R. Vassar, K. Cahill-Rowley, K. W. Yeom, D. K. Stevenson, and J. Rose, "Prediction of cognitive and motor development in preterm children using exhaustive feature selection and cross-validation of near-term white matter microstructure," *NeuroImage: Clinical*, vol. 17, pp. 667–679, 2018.

[2] K. Cahill-Rowley, K. Schadl, R. Vassar, K. W. Yeom, D. K. Stevenson, and J. Rose, "Prediction of gait impairment in toddlers born preterm from near-term brain microstructure assessed with dti, using exhaustive feature selection and cross-validation," *Frontiers in Human Neuroscience*, vol. 13, p. 305, 2019.

[3] R. Vassar, K. Schadl, K. Cahill-Rowley, K. Yeom, D. Stevenson, and J. Rose, "Neonatal brain microstructure and machine-learning-based prediction of early language development in children born very preterm," *Pediatric Neurology*, vol. 108, pp. 86–92, 2020.

[4] N. J. Stevenson, L. Oberdorfer, N. Koolen, J. M. O'Toole, T. Werther, K. Klebermass-Schrehof, and S. Vanhatalo, "Functional maturation in preterm infants measured by serial recording of cortical activity," *Scientific Reports*, vol. 7, no. 1, p. 12969, 2017.

[5] N. J. Stevenson, L. Oberdorfer, M.-L. Tataranno, M. Breakspear, P. B. Colditz, L. S. de Vries, M. J. N. L. Benders, K. Klebermass-Schrehof, S. Vanhatalo, and J. A. Roberts, "Automated cot-side tracking of functional brain age in preterm infants," *Annals of Clinical and Translational Neurology*, vol. 7, no. 6, pp. 891–902, 2020.

[6] H. J. McCrea and L. R. Ment, "The diagnosis, management, and postnatal prevention of intraventricular hemorrhage in the preterm neonate," *Clinics in Perinatology*, vol. 35, no. 4, pp. 777–792, 2008.

[7] W. Deng, J. Pleasure, and D. Pleasure, "Progress in periventricular leukomalacia." *Arch Neurol*, vol. 65, no. 10, pp. 1291–1295, Oct 2008.

[8] R. D. Higgins, A. H. Jobe, M. Koso-Thomas, E. Bancalari, R. M. Viscardi, T. V. Hartert, R. M. Ryan, S. G. Kallapur, R. H. Steinhorn, G. G. Konduri, S. D. Davis, B. Thebaud, R. I. Clyman, J. M. Collaco, C. R. Martin, J. C. Woods, N. N. Finer, and T. N. K. Raju, "Bronchopulmonary dysplasia: Executive summary of a workshop." *Journal of Pediatrics*, vol. 197, pp. 300–308, Jun 2018.

[9] W. H. Yee, A. S. Soraisham, V. S. Shah, K. Aziz, W. Yoon, S. K. Lee, and , "Incidence and timing of presentation of necrotizing enterocolitis in preterm infants," *Pediatrics*, vol. 129, no. 2, pp. e298–e304, 2012.

[10] S. Bolisetty, A. Dhawan, M. Abdel-Latif, B. Bajuk, J. Stack, J.-L. Oei, K. Lui, and , "Intraventricular hemorrhage and neurodevelopmental outcomes in extreme preterm infants," *Pediatrics*, vol. 133, no. 1, pp. 55–62, 2014.

[11] T. Imamura, H. Ariga, M. Kaneko, M. Watanabe, Y. Shibukawa, Y. Fukuda, K. Nagasawa, A. Goto, and T. Fujiki, "Neurodevelopmental outcomes of children with periventricular leukomalacia," *Pediatrics & Neonatology*, vol. 54, no. 6, pp. 367–372, 2013.

[12] J. Y. Choi, D.-w. Rha, and E. S. Park, "The effects of the severity of periventricular leukomalacia on the neuropsychological outcomes of preterm children," *Journal of Child Neurology*, vol. 31, no. 5, pp. 603–612, 2021/04/29 2015.

[13] P. J. Anderson and L. W. Doyle, "Neurodevelopmental outcome of bronchopulmonary dysplasia," *Seminars in Perinatology*, vol. 30, no. 4, pp. 227–232, 2006.

[14] S. R. Hintz, D. E. Kendrick, B. J. Stoll, B. R. Vohr, A. A. Fanaroff, E. F. Donovan, W. K. Poole, M. L. Blakely, L. Wright, and R. Higgins, "Neurodevelopmental and growth outcomes of extremely low birth weight infants after necrotizing enterocolitis," *Pediatrics*, vol. 115, no. 3, pp. 696–703, 2005.

[15] C. M. Rees, A. Pierro, and S. Eaton, "Neurodevelopmental outcomes of neonates with medically and surgically treated necrotizing enterocolitis," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 92, no. 3, pp. F193–F198, 2007.

[16] C. R. Martin, O. Dammann, E. N. Allred, S. Patel, T. M. O'Shea, K. C. Kuban, and A. Leviton, "Neurodevelopment of extremely preterm infants who had necrotizing enterocolitis with or without late bacteremia," *The Journal of Pediatrics*, vol. 157, no. 5, pp. 751–756.e1, 2010.

[17] J.-M. Moutquin, "Classification and heterogeneity of preterm birth," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 110, no. s20, pp. 30–33, 2003.

[18] "Committee opinion no 579: Definition of term pregnancy," *Obstetrics & Gynecology*, vol. 122, no. 5, 2013.

[19] "Age terminology during the perinatal period," *Pediatrics*, vol. 114, no. 5, pp. 1362–1364, 2004.

[20] J. Z. Bakdash and L. R. Marusich, "Repeated measures correlation," *Frontiers in Psychology*, vol. 8, p. 456, 2017.

[21] "rmcorr: Repeated measures correlation," CRAN, Jan. 2021. [Online]. Available: https://CRAN.R-project.org/package=rmcorr

Conclusion and Future Work

In this dissertation we focused on proposing decision support systems (DSS), using machine learning techniques applied to physiological signals, for two main goals: the early diagnosis of late-onset sepsis (LOS), and the objective evaluation of maturation in preterm infants hospitalized in neonatal intensive care unit (NICU). The DSSs proposed should be a reliable and non-invasive tool, to help physicians better assess, in real time, the situation of the neonates in order to take key decisions in a timely manner regarding the course of treatment and care of the infants.

LOS is one of the leading causes of morbidity and mortality among preterm infants ([1, 2]). Therefore, it is not surprising that multiple studies have focused on the development of machine learning models for its early LOS detection. These previous efforts have ranged in their input data from using physiological and clinical data, including results of laboratory tests [3], to focusing exclusively on heart rate characteristics [4].

The work presented in this thesis was framed in the Digi-NewB project, which aims to combine clinical signs, physiological signals, and video and sound recordings in DSS for neonatal monitoring. However, and although we acknowledge the importance of clinical features and laboratory tests results and their potential contribution to a DSS, we decided to focus on using exclusively heart rate variability (HRV) data. This decision was based on the fact that this relies exclusively on the heart rate signal, which is routinely and continuously monitored in NICU, so it does not require any invasive tests of additional equipment to be in contact with the infant. Also, it does not require the health care personnel to manually input any data, nor to wait for laboratory test results. These features would allow our proposed DSS to be a real-time and non-invasive tool, that can still be robust and reliable, for the early detection of LOS. Instead, we focused on improving the methods which have been proposed for LOS detection that rely on heart rate data. We took a two way approach to doing this: first, by studying the introduction of new HRV features that could improve the performance of the models; second, by testing more powerful models that can better capture complex patterns in the data and take into account the evolution of the infection over time.

The objective of maturation in preterm infants, in comparison, has been less explored in previous literature. While some studies have focused in evaluating the maturation using magnetic resonance (MRI) [5] imaging or electroencephalogram (EEG) data [6], we did not find any previous studies that aimed to evaluate the maturation of preterm neonates based on heart rate data.

The previous studies in this field have also taken different approaches. With a series of studies focusing on predicting which prematurely born infants would develop cognitive [7], motor [8], or language [9] impairment by the time they reached two years of age, using MRI data from

the perinatal period. Other studies, instead, focused on using either MRI [10] or EEG [6] to estimate a functional maturational age. In this dissertation we took an approach more similar to the later, using physiological signals to estimate a functional maturational age (FMA), under the hypothesis that a larger difference between this FMA and the postmesntrual age (PMA), which is clinically determined, can indicate an abnormal maturation and guide both short-term and long-term management and intervention.

Furthermore, although the machine learning method we developed for the maturation study was also initially designed for HRV data, we were able to proof that the same model generalizes well to other types of data available as part of the Digi-NewB project, in which our work was framed. In particular, we tested the model with respiration rate variability (RRV) and Bradycardia data, obtaining positive results.

Summary of Findings Regarding Late-Onset Sepsis Diagnosis

For the goal of improving LOS detection systems, we argued that the inclusion of visibility graph indexes, which are derived from a network-analysis of the HRV time-series, would improve the performance of the machine learning algorithms. This hypothesis was based on previous studies which suggested that these features added different information about the HRV than time and frequency domain features and non-linear measurements ([11, 12], as well as previous study which found that the inclusion of visibility graph features improved sepsis detection in adults [13].

To test this hypothesis we trained four different types of machine learning algorithms (namely, k-nearest neighbours, logistic regression, random forest, and support vector machines) on datasets derived from the same population, with one version of the feature set including the visibility graph features and the other one excluding them. We compared the results from all the machine learning algorithms, and found that in three out of the four machine learning algorithms, the performance was improved when visibility graph indexes were included in the feature set. The best performing algorithm was the logistic regression when visibility graph indexes were part of the feature set, achieving an area under the receiver operating characteristics curve (AUROC) of 87.7%. In fact, the introduction of these features increased the performance of this algorithm by 6.8% in terms of its AUROC, and the likelihood ratio analysis suggested that the improvement introduced by these features is statistically significant, with p-value $< 3e{-}4$. Based on these findings, we included the visibility graph indexes in the feature sets used for all posterior studies based on HRV which were discussed in this dissertation, including those focused on the study of maturation.

Another important contribution of this study was regarding the search for optimal calibration period and learning window. We introduced the idea of a calibration period, based on a

similar approach from previous studies [14], to account for the differences between the HRV characteristics of infants with different degrees of prematurity, as well as for the changes caused by the LOS itself. We found that the optimal calibration period was of 48 hours for three out of the four machine learning algorithms tested, including the best performing one, which was the logistic regression model.

Similarly, we tested different learning windows, by varying how many hours before the diagnosis of LOS were labeled as infected for the patients in the LOS group. This is an important study, given that, with the current medical knowledge and technology, it is impossible to clinically determine when exactly a patient became infected, which translates into difficulties when training machine learning algorithm as it likewise impossible to determine with certainty which samples from septic patients to label as LOS and which as healthy. In this regard, we found more divergence in our results, however, the best performing algorithm used a learning window of 42 hours before LOS diagnosis. This means, that the 42 hours before clinically diagnosis were labeled as LOS. While the other algorithms had better results, in terms of their AUROC, with different learning windows, all of them still achieved some of their highest performance for the 42 hours learning window. These findings suggests that 42 before clinical diagnosis of sepsis might be, on average, a good and evidence-based approximation for the period to label as LOS in future studies.

Therefore, in the second study presented in this dissertation regarding LOS diagnosis, we used a learning window of 42 hours, labeling all hours before this period as not LOS for the patients who were in the LOS group. In said study we aimed to test the use of recursive neural networks (RNN) for early LOS diagnosis. This was inspired by the fact that most current studies focused on the application of machine learning for the objective of sepsis diagnosis in preterm infants have relied on fairly simple models, such as logistic regression [4]. RNNs, which have been used successfully applied to sepsis diagnosis in adults, are more powerful, and can detect complex time patterns in the data due its recurrent connections which allow it to have memory [15]. This makes RNN models specially well suited for studying time series, which is the data type that concerns us, whether it is the time series of HRV features, or the raw HRV time series.

In fact, for this study we trained RNNs models on both the raw HRV time series and the time series of HRV features. While the performance was relatively low in the first case, we showed that the model is indeed capable of learning and therefore suggest further studies with more data, as a model based on raw HRV time series as the one we proposed, would have the advantage of offering a more continuous evaluation while requiring less preprocessing. However, the RNN based on the HRV features achieved a very high performance, with an AUROC consistently above 80% for the 24 hours before clinical diagnosis of LOS, and reaching a maximum of 90.4% for the period of six hours before LOS diagnosis. Besides its remarkable performance, this model already offers some advantages over more simple models such as the ones we suggested in the

first study, as it can capture time patterns and changes without requiring a calibration period and the extensive feature engineering associated with it.

Our findings serve as a proof of concept for a non-invasive and real-time DSS based on HRV data, including visibility graph analysis, and the use of RNN for the early diagnosis of LOS in preterm infants. Furthermore, with the help of other partners of the Digi-NewB project, an interface was conceived to follow the time evolution of the LOS probability indices we proposed.

Summary of Findings Regarding the Evaluation of Maturation

For the evaluation of maturation we defined as the metric a functional maturational age, which is the output of a machine learning model trained with the postmenstrual age as target variable. We argued that the FMA would be closer to the PMA in preterm infants who did not show signs of delayed maturation, nor were diagnosed with medical conditions associated with negative neurodevelopmental outcomes during their hospitalization in NICU. While a bigger difference between FMA and PMA would be associated with a disruption in the maturation.

To test this, we designed and developed a machine learning method that involves automated feature selection and ensemble machine learning. For the feature selection we proposed the application of a filtering step, based on eliminating features from the dataset that were very weekly correlated to the PMA. This step was then followed by the application of a genetic algorithm on the remaining features, to choose the final optimal set of features to be used by the ensemble machine learning algorithm. Such algorithm consisted of a linear regression model, which output is then passed as an additional feature to a random forest regression model. Given that the ensemble model combines two different algorithms, two instances of the genetic algorithm were used, to find the optimal feature set for each component of the machine learning model.

The model was initially designed to work with HRV data as input, and the application of the two instances of the genetic algorithm for feature selection in the described manner yielded interesting insights regarding the behaviour of the HRV features in relation to the PMA. The feature selection process revealed a preference of linear regression for time-domain and visibility graph features, suggesting a linear correlation between these features and the PMA. Instead, the random forest regression favored frequency-domain features, which suggests a non-linear correlation between these features and PMA.

At a later stage of the development phase, the model was generalized to function with different data types. Thus, the final model was tested on three different data types available in the Digi-NewB project at the time of the study, which were, in addition to the HRV, respiration rate variability (RRV) and bradycardia data. This resulted in three instances of the model being trained, one on each data type. At first we used a population of only healthy infants to train

and test the model. Thus, we verified that for healthy preterm infants, with varying degrees of prematurity, but further no signs of delayed maturation during NICU hospitalization, the model could estimate an FMA that was very close to the infants' PMA, with an average mean absolute error (MAE) 0.93 weeks with the HRV data, and of 1.39 weeks with RRV and bradycardia data.

Afterwards, we tested our proposed method on a population of infants diagnosed in the perinatal period with medical conditions associated with negative effects on the neurodevelopment of premature infants, and compared the results to those obtained on a healthy population. To this end, we trained the three instances of the proposed model, one for each data type, using a population of only healthy preterm individuals. Then, we used those trained models to make predictions in a test set of healthy preterm infants, as well as on a population of neonates diagnosed with medical conditions known to affect neurodevelopmental maturration. Specifically, the later population was constituted by infants who were diagnosed with either neurological lesions (NL), bronchopulmonary dysplasia (BDP), necrotizing enterocolitis (NEC), or a combination of two or more of these conditions.

For all three data types we found that the healthy test patients presented the lowest MAE, as well as the highest correlation between FMA and PMA, in comparison to infants in the population with compromised maturation. Furthermore, we found that the model seemed to be more sensitive to maturational disruptions in the infants diagnosed with NL, NEC, or with multiple conditions, than to infants with only BPD. In fact, from the population of infants with abnormal maturation, those with only BPD presented the highest correlation between FMA and PMA for all three data types, and the lowest MAE for the HRV and RRV data. Finally, we also showed a graphical representation of the results in a spider chart, where we describe the evolution of the FMA along different axis, each one corresponding to one data type.

These results suggest that the ensemble machine learning we proposed can indeed estimate a functional maturational age based on different data types obtained routinely in NICU. We consider that the evidence presented in this dissertation in regards to the evaluation of maturation serves as a proof of concept for a DSS based on the estimation of the FMA as an objective metric of the maturation progress of the infants during their hospitalization. Such a metric could help physicians in decision making regarding short-term decisions, such as when to terminate ventilatory support or to discharge a patient from the NICU, as well as long-term decisions to help improve neurodevelopmental outcomes.

Strengths and Limitations

Legal and ethical considerations of working with neonates pose a difficulty for the data acquisition from this type of population. And even once these obstacles are surpassed, it requires the collaboration of the medical staff and parental agreement. Finally, there are factors which

are out of the researchers and medical personnel control, such as which babies will present or not the conditions to make them eligible for the studies. This results in a database which is considered small in the field of machine learning. Therefore, it would be necessary to acquire an even larger database, specially including more infants who developed LOS, in order to minimize the data unbalance between groups in the sepsis study, for which we had a ratio of approximately 5 infants in the control group for every infant in the LOS group. The maturation study could also benefit from a larger database, including both more healthy and unhealthy infants. This would allow us to train more reliable and robust models.

However, one strong point of this dissertation is that all these data were acquired as part of a dedicated study, thus ensuring that all data acquisition materials and methods, as well as the signal processing, are the same for all patients in the population. Also, the final decisions for inclusion or exclusion from the population were taken by the same physicians. From a machine learning and statistical analysis point of view, this is an advantage, as it minimizes potential differences between the patients that are related to different acquisition methods rather than actual physiological differences between the patients.

Another limitation of our work is that we did not include clinical data or laboratory test results in our models. We acknowledge that these type of data can introduce additional and relevant information to machine learning models both for LOS diagnosis and maturation evaluation. However, we chose to exclude such features because we did not want our models to rely on invasive techniques (as laboratory tests would be, for example), nor on information that needs to be manually entered by medical personnel, as this can inconvenience them, and it is also more prone to human error, because it can be forgotten or wrongly annotated on occasions. Moreover, these type of data are not continuous, which would also require additional data preprocessing steps to allow for a DSS to work in real-time. For these reasons, the only clinical data we used was the gestational age for the maturation study. However, as the gestational age is a constant, it only needs to be annotated once, at the beginning of hospitalization, so it does not pose the same inconveniences as other clinical data.

Nonetheless, the fact that we only use data extracted from the heart rate monitoring (with the aforementioned exception of the gestational age), which is routinely and continuously done in NICU allows the proposed DSS, both for LOS detection and maturation evaluation, to be non-invasive, to require negligible additional effort from health care personnel, and to work in nearly real-time, while still achieving a high performance. We consider that these characteristics are a strength of the DSS proposed in this dissertation.

Another noteworthy element of our work is that, throughout this dissertation and for every proposed system, we showed sample cases of what the output of the model looks like for individual patients, rather than focusing exclusively on global metrics such as the AUROC, MAE, or correlation between variables. This allowed us to exemplify how the DSS could be used in

real life, and how their final output could be presented in a manner that is easy to read and interpret by medical staff. Thus, we offer a proof of concept for the DSSs, not only in terms of high global performance but also of their usability in a patient specific manner.

Finally, we consider that one of the strongest points of our research, specifically concerning the maturation study, is that we were able to design and develop a method that can be generalized to all data types included in the Digi-NewB project. This includes not only HRV, RRV, and bradycardia, which were tested and presented in this dissertation, but also movement, cry, and sleep data. It is our hope that this will facilitate future research in the Digi-NewB framework, and possibly in other projects as well.

Future Directions

There are several steps to take to expand the reach of the present work in the short-term. One of the first and most urgent ones must be to validate the methods and decision support systems proposed in this dissertation, both for LOS diagnosis and maturation evaluation, in larger databases, possibly even including data external to the Digi-NewB project. This would allow for the training of more robust and reliable model that are better able to generalize when faced with previously unseen data.

For the sepsis study, we consider that the next step should be focused on further developing a model based on RNN and using the raw HRV time series as input. Possible directions might include compressing the signal, or computing the time series spectrogram and then process it as an image. On the other hand, with a large enough database, an RNN model could be tested even in the raw electrocardiogram data. In fact, a machine learning approach for LOS detection, based on artificial neural networks and raw physiological data has already been proposed by the Digi-NewB partners in Tampere University, Finland, in the last project report.

Another interesting next step would be to include RRV and bradycardia data for early LOS detection. This could be done by either integrating these into one joint dataset along with the HRV features, or by producing a different score for each data type, in a similar manner to the approach we took on the maturation study. However, any such developments should be done in close collaboration with physicians, in order to offer not only a technical interpretation of the results from a machine learning perspective, but also an interpretation of the underlying physiological mechanisms leading to such results.

Regarding the maturation study, the next step should be testing the proposed approach on the features extracted from movement, cry, and sleep data, once those datasets are processed and stabilized. It would also be interesting to test this method using data external to the Digi-NewB project. For instance, this could be done using HRV data, but described in terms of different features to the ones we used; or even completely different data which also has a correlation with

maturation, such as features derived from EEG recordings. This would allow us to examine not only how well the method generalizes, but also its user-friendliness for researchers who are not familiar with the Digi-NewB project.

For the maturation study it would also be of great interest to do a retrospective analysis to determine if the premature infants whose FMA presented more deviation from the PMA display any signs of delayed maturation later in life. Such a study would also allow us to analyze if there is a correlation between how big difference between the FMA and PMA is during the postnatal period, and the degree or type neurodevelopmental compromise present later in life.

Finally, in the long-term we envision testing both, the LOS DSS and the maturation evaluation DSS in dedicated clinical trials. This would allow us to evaluate if they are user-friendly, acceptable, and easily interpretable for physicians and nurses in NICU, and to determine the feasibility of deploying in a real-life and real-time NICU scenario. But most importantly, such a study would indicate if these DSS can have an impact reducing in morbidity, mortality, and length of hospitalization for preterm infants.

Bibliography

[1] B. J. Stoll, N. Hansen, A. A. Fanaroff, L. L. Wright, W. A. Carlo, R. A. Ehrenkranz, J. A. Lemons, E. F. Donovan, J. E. T. Ann R. Stark, W. Oh., C. R. Bauer, S. B. Korones, S. Shankaran, A. R. Laptook, D. K. Stevenson, L.-A. Papile, and W. K. Poole, "Late-Onset Sepsis in Very Low Birth Weight Neonates: The Experience of the NICHD Neonatal Research Network," *Pediatrics*, vol. 110, no. 2, pp. 285–291, Aug. 2002.

[2] L. Liu, H. L. Johnson, S. Cousens, J. Perin, J. E. L. Susana Scott, I. Rudan, H. Campbell, R. Cibulskis, M. Li, R. E. B. Colin Mathers, for the Child Health Epidemiology Reference Group of WHO, and UNICEF, "Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000," *The Lancet*, vol. 379, no. 9832, pp. 2151–2161, Jun. 2012.

[3] S. Mani, A. Ozdas, C. Aliferis, H. A. Varol, Q. Chen, R. Carnevale, Y. Chen, J. Romano-Keeler, H. Nian, and J.-H. Weitkamp, "Medical decision support using machine learning for early detection of late-onset neonatal sepsis," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 326–336, 09 2013.

[4] M. P. Griffin, T. M. O'Shea, E. A. Bissonette, F. E. Harrell, D. E. Lake, and J. R. Moorman, "Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness," *Pediatric Research*, vol. 53, no. 6, pp. 920–926, 2003.

[5] C. D. Smyser, N. U. F. Dosenbach, T. A. Smyser, A. Z. Snyder, C. E. Rogers, T. E. Inder, B. L. Schlaggar, and J. J. Neil, "Prediction of brain maturity in infants using machine-learning algorithms," *NeuroImage*, vol. 136, pp. 1–9, 08 2016.

[6] N. J. Stevenson, L. Oberdorfer, N. Koolen, J. M. O'Toole, T. Werther, K. Klebermass-Schrehof, and S. Vanhatalo, "Functional maturation in preterm infants measured by serial recording of cortical activity," *Scientific Reports*, vol. 7, no. 1, p. 12969, 2017.

[7] K. Schadl, R. Vassar, K. Cahill-Rowley, K. W. Yeom, D. K. Stevenson, and J. Rose, "Prediction of cognitive and motor development in preterm children using exhaustive feature selection and cross-validation of near-term white matter microstructure," *NeuroImage: Clinical*, vol. 17, pp. 667–679, 2018.

[8] K. Cahill-Rowley, K. Schadl, R. Vassar, K. W. Yeom, D. K. Stevenson, and J. Rose, "Prediction of gait impairment in toddlers born preterm from near-term brain microstructure assessed with dti, using exhaustive feature selection and cross-validation," *Frontiers in Human Neuroscience*, vol. 13, p. 305, 2019.

[9] R. Vassar, K. Schadl, K. Cahill-Rowley, K. Yeom, D. Stevenson, and J. Rose, "Neonatal brain microstructure and machine-learning-based prediction of early language development in children born very preterm," *Pediatric Neurology*, vol. 108, pp. 86–92, 2020.

[10] P. Galdi, M. Blesa, D. Q. Stoye, G. Sullivan, G. J. Lamb, A. J. Quigley, M. J. Thrippleton, M. E. Bastin, and J. P. Boardman, "Neonatal morphometric similarity mapping for predicting brain age and characterizing neuroanatomic variation associated with preterm birth," *NeuroImage: Clinical*, vol. 25, p. 102195, 2020.

[11] T. Madl, "Network analysis of heart beat intervals using horizontal visibility graphs," *2016 Computing in Cardiology Conference (CinC)*, pp. 733–736, 2016.

[12] T. Nguyen Phuc Thu, A. I. Hernández, N. Costet, H. Patural, V. Pichot, G. Carrault, and A. Beuchée, "Improving methodology in heart rate variability analysis for the premature infants: Impact of the time length," *PLOS ONE*, vol. 14, no. 8, pp. 1–14, 08 2019.

[13] S. P. Shashikumar, Q. Li, G. D. Clifford, and S. Nemati, "Multiscale Network Representation of Physiological Time Series for Early Prediction of Sepsis," *Physiological Measurement*, vol. 38, no. 12, pp. 2235–2248, Nov. 2017.

[14] L. B. Mithal, R. Yogev, H. L. Palac, D. Kaminsky, I. Gur, and K. K. Mestan, "Vital signs analysis algorithm detects inflammatory response in premature infants with late onset sepsis and necrotizing enterocolitis," *Early Human Development*, vol. 117, pp. 83–89, 2018.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," Dec. 2014. [Online]. Available: https://arxiv.org/abs/1412.3555v1

List of publications

International Journals

1. **C. León**, G. Carrault, P. Pladys, and A. Beuchée, "Early detection of late onset sepsis in premature infants using visibility graph analysis of heart rate variability," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1006–1017, 2021.
2. **C. León**, S. Cabon, H. Patural, G. Gascoin, C. Flamant, J.-M. Roué, G. Favrais, A. Beuchée, P. Pladys, and G. Carrault, "Evaluation of maturation in preterm infants through an ensemble machine learning algorithm using physiological signals," *IEEE Journal of Biomedical and Health Informatics*, 29 June 2021.
3. G. Bury, S. Leroux, **C. León Borrego**, C. Gras Leguen, D. Mitanchez, G. Gascoin, A. Thollot, J. M. Roué, G. Carrault, P. Pladys, and A. Beuchée, "Diagnosis of neonatal late-onset infection in very preterm infant: Inter-observer agreement and international classifications," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, 2021.

International Conferences

1. **C. León Borrego**, G. Carrault, C. Flamant, J.-M. Roué, G. Gascoin, G. Favrais, A. Thollot, A. Hernandez, A. Beuchée, P. Pladys, and Digi-NewB consortium, "Early detection of late onset sepsis in preterm infants using machine learning and visibility graph for analysis of heart rate variability," in *European Congress on Perinatal Medicine 2020*, 2020.
2. P. Pladys, **C. León Borrego**, G. Carrault, C. Flamant, G. Gascoin, J.-M. Roué, G. Favrais, A. Thollot, A. Beuchée, and A. Hernandez, "Early detection of late onset sepsis in preterm infants using machine learning and visibility graph for analysis of heart rate variability," in *European Academy of Paediatric Societies 2020*, 2020.

Software

1. S. Cabon, **C. León Borrego**, F. Porée, G. Carrault, PREDICT : Prediction d'un Identificateur Cible à partir d'un ensemble de Paramètres

UNIVERSITÉ DE RENNES 1

Titre : Apprentissage automatique pour la prédiction de l'infection et de la maturation chez le grand prématuré en associant les variabilités cardiaques et respiratoires.

Mot clés : Réseaux neuronaux récurrents, apprentissage automatique d'ensemble, système d'aide à la décision, graphes de décision, signaux physiologiques, variabilité cardiaque, infection, maturation.

Résumé : Cette thèse s'inscrit dans le cadre du projet européen Digi-NewB dont l'objectif principal était de développer un nouveau système de surveillance des grands prématurés. Le projet a impliqué des partenaires de quatre pays et a permis de collecter des données cliniques, des signaux physiologiques, des données vidéo et des pleurs de bébés dans six hôpitaux en France.

L'objectif plus spécifique de ce travail était de proposer des systèmes d'aide à la décision (DSS) basés sur des modèles d'apprentissage automatique pour le diagnostic précoce de l'infection tardive (LOS) et pour l'évaluation de la maturation des prématurés. Parmi les différentes données acquises dans le cadre du projet, nous nous sommes concentrés sur la variabilité de la fréquence cardiaque (HRV), la variabilité de la fréquence respiratoire (RRV) et les données de bradycardie.

Les principales contributions de ce travail ont été : (i) l'intérêt des indices mesurés sur les graphes de visibilité caractérisant la HRV pour la détection du LOS ; (ii) la proposition d'un réseau neuronal récurent performant pour le diagnostic précoce du LOS fondé sur les paramètres de la HRV ; (iii) l'introduction d'un modèle d'apprentissage ensembliste pour le suivi de la maturation des enfants prématurés à partir de la HRV, de la RRV ou des bradycardies ; (iv) la preuve de concept de ce modèle sur une population comprenant des enfants prématurés avec une maturation normale et anormale. Il importe aussi de souligner que tous ces développements ont été effectués dans un souci d'exploitation en temps réel et que la preuve de faisabilité que cela soit pour le LOS ou la maturation a été effectuée en unité de soins intensifs néonatale.

Title: Machine learning for the prediction of infection and evaluation of maturation in premature infants combining cardiac and respiratory variability.

Keywords: Recurrent neural networks, ensemble machine learning, decision support system, physiological signals, infection, maturation

Abstract: This dissertation was framed in the Digi-NewB project, which was founded by the European Union, and had as main goal to improve health care for neonates through the development of new monitoring systems. The project involved partners from four countries, and collected health records, physiological signals, and video and sound data from infants in six hospitals in France.

The objective of our research was to propose decision support systems (DSSs) based on machine learning models for the early diagnosis of LOS and for the evaluation of maturation in preterm infants. From the data types acquired in the project, we limited our scope to physiological signals. We focused on heart rate variability (HRV), respiration rate variability (RRV), and bradycardia data.

The main contributions of this work are: (i) an assessment of the positive impact in the performance of machine learning models for the detection of LOS of including visibility graph indexes for the characterization of HRV; (ii) a high performing recursive neural network model for early diagnosis of LOS in preterm infants based on HRV features; (iii) an ensemble machine learning model for the evaluation of maturation of the infants in terms of their functional maturational age, derived from HRV, RRV, or bradycardia features; (iv) the validation of this model on a population that included preterm infants with normal and abnormal maturation. The models presented in this work serve as proof of concept for non-invasive DSSs that can have a high performance in real-time.