Development of Machine Learning Models for Hemophilia A Severity Classification Based on FVIII Protein Mutations

1st Vitor Bonella
Department of Informatics
Federal University of Espírito Santo
Vitória, Brazil
vitor.bonella@edu.ufes.br

***Abstract*—Hemophilia A is a rare hereditary bleeding disorder caused by a deficiency in clotting factor VIII (FVIII), leading to prolonged bleeding episodes and affecting approximately 1 in 5,000 male births globally. Traditional diagnostic methods involve invasive blood tests to measure FVIII levels, which pose significant risks for patients with severe Hemophilia A. This study aims to develop machine learning classification models to distinguish between mild, moderate, and severe Hemophilia A using data on point mutations in the FVIII protein. The dataset comprises 443 records of point mutations, including genetic attributes, protein structural features, and interaction network data. After data cleansing and standardization, the refined dataset contains 415 records. Multiple machine learning algorithms were evaluated, with the Random Forest model emerging as the most effective in classifying Hemophilia A severity. Despite its promising results, further enhancement is necessary for clinical application. Expanding the dataset and incorporating advanced machine learning techniques will be crucial for improving model performance and reliability. This research marks a significant step forward in non-invasive diagnostic approaches for Hemophilia A and contributes to the broader application of AI in rare disease management.**

***Index Terms*—Hemophilia A, Machine learning, Diagnostic methods**

I. Introduction

Hemophilia A is a hereditary bleeding disorder characterized by a deficiency in clotting factor VIII (FVIII), leading to prolonged bleeding episodes. This condition predominantly affects males, with an incidence of approximately 1 in 5,000 male births worldwide [13]. The management and treatment of Hemophilia A have evolved significantly over the years, yet early and accurate diagnosis remains a critical challenge.

Recent advances in artificial intelligence (AI) have shown significant potential in revolutionizing the healthcare sector, offering new avenues for the diagnosis, treatment, and management of various diseases, including hemophilia A. AI techniques, particularly machine learning (ML), have demonstrated remarkable capabilities in analyzing complex medical data and generating predictive models that surpass traditional statistical methods [14].

Hemophilia presents a unique challenge in the medical landscape due to its rarity compared to more prevalent conditions like heart disease, diabetes, and depression. With only about 250,000 affected individuals globally, as noted by Hemophilia [7], there is a notable scarcity of cases. This scarcity translates into limited financial investment and research efforts aimed at understanding and managing the condition effectively. In light of this scenario, this project proposes leveraging AI, specifically machine learning techniques, to develop an alternative diagnostic approach for assessing the severity of type A hemophilia. This endeavor builds upon the pioneering work discussed by Lopes et al. [8], which established a correlation between mutations in FVIII and the predictive assessment of

Hemophilia A severity.

This project endeavors to create classification models capable of distinguishing between the three severities of Hemophilia A (mild, moderate, and severe) by leveraging data on point mutations in the FVIII protein and its structural and functional characteristics. This data can be sourced from DNA sequencing, potentially obtained from nasal swab samples. This approach aims to replace invasive standard blood collection tests for measuring FVIII levels, which can pose risks for patients with severe Hemophilia A due to difficulties in blood clotting during collection. The model results were thoroughly compared, ultimately identifying Random Forest as the most effective final model. This machine learning endeavor significantly enhances our understanding and exploration of Hemophilia A severity and its associated FVIII protein, marking a notable advancement in the research and comprehension of this rare genetic condition.

The remainder of this paper is structured as follows: Section II offers a contextual overview of the problem and outlines previously applied methods. Section III elaborates on the image descriptors extracted from the public lighting dataset's digital images. Section IV delves into the techniques utilized in detail. Section V outlines the experimental methodology, dataset details, classifier selection, and feature selection process. Section VI presents the experimental outcomes along with their analysis. Finally, Section VII concludes with findings and outlines future research directions.

II. Hemophilia Classification

Data from the World Federation of Hemophilia (WFH) annual report indicates that around 250,000 individuals worldwide are afflicted with hemophilia, rendering it a rare genetic disorder. Hemophilia, which is linked to the X chromosome, predominantly impacts males due to genetic inheritance. Females, possessing two X chromosomes, seldom develop the ailment unless there are rare occurrences of mutations in both chromosomes. This condition compromises the clotting ability of blood and can stem from either inherited mutations (70% of cases) or spontaneous mutations (30% of cases),

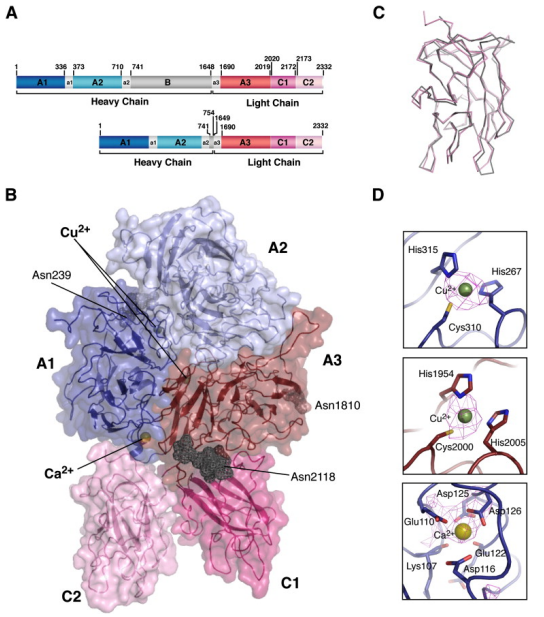

Fig. 1. The structure of FVIII, illustrated by Ngo et al. [11], consists of 2332 amino acids grouped into distinct domains: A1, A2, A3, B (not depicted), C1, and C2. Among these, the A, B, and C domains are pivotal for the protein's structure and function. The A domain is particularly essential for FVIII activation, while the B domain, although not shown in the figure and less comprehensively understood, is the largest among the primary domains. Lastly, the C domain is believed to impact another critical factor, von Willebrand factor, which contributes to the blood coagulation cascade [9].

which may not necessarily have a familial precedent [12].

There exist two primary types of hemophilia: Type A, characterized by a deficiency in clotting factor VIII (FVIII), and Type B, associated with a deficiency in clotting factor IX (FIX)[4]. Hemophilia A constitutes the majority of cases, accounting for approximately 80 to 85% of incidents, while Hemophilia B encompasses the remaining cases. These statistics align with the WFH report, which indicated in 2020 that around 80% of patients had Hemophilia A, approximately 15% had Hemophilia B, and the remaining 5% were unidentified.

The classification of hemophilia severity is contingent upon the level of clotting factor activity, categorizing cases as mild, moderate, or severe.

Severe cases present substantial risks of potentially fatal spontaneous bleeding, necessitating continuous care and specialized treatment. The conventional treatment for hemophilia involves the replacement of deficient clotting factors, with the specific approach varying according to the severity of the cases. However, the considerable costs associated with treatment, particularly for severe cases, underscore the significance of accurate and efficient diagnosis [6].

Diagnosing hemophilia entails a comprehensive analysis of family history, symptoms, and laboratory tests to assess clotting factor levels. Nonetheless, it is imperative to enhance this diagnostic process by exploring less invasive and safer methods, particularly for severe cases. The adoption of alternative techniques, such as substituting invasive tests with less intrusive methodologies to diagnose hemophilia severity, can mitigate complications in severe patients [1].

Given that the majority of hemophilia cases pertain to Type A, it is paramount to devise innovative and less invasive diagnostic strategies to identify and evaluate the severity of this form of the disease. Achieving this objective necessitates a detailed exploration of the FVIII protein (Figure 1), as it plays a pivotal role in the etiology of Hemophilia A.

III. Dataset Description

The Hemophilia dataset, comprising 443 records, each denotes a distinct point mutation within the FVIII protein, delineating the precise location of the mutation along with the amino acid sequence before and after the mutation event. Among the 20 attributes, the dataset encapsulates genetic attributes pertinent to the mutation, structural features of the protein at the mutation site, and insights into the interaction network among residues within this protein. Additionally, a variable characterizing the severity of Hemophilia A is included.

The compilation of this database was executed through a meticulous two-step process. Initially, a thorough examination and cleansing of null data resulted in the elimination of 28 records. Subsequently, textual data underwent standardization measures, which involved the removal of redundant white spaces and the establishment of uniform representation of letters to ensure precise identification and differentiation of categorical variables. Upon the culmination of these procedures, the dataset underwent a reduction in size, concluding with a refined collection of 415 records.

- **AAHGVS (numeric):** Denotes the position of the amino acid where the mutation transpired, considering the complete conformation of the protein.
- **AALegacy (numeric):** Represents the position of the amino acid where the mutation occurred, accounting for the mature state of the protein.
- **aaOne (categorical):** Refers to the original amino acid preceding the mutation event.
- **AAdist (numeric):** Quantifies the distance between the original amino acid (aaOne) and the resultant amino acid after mutation [8].
- **Domain (categorical):** Signifies the specific domain within the protein where the mutation manifested.
- **psi (numeric):** Describes the torsion angle around the $C_\alpha - C$ bond.
- **phi (numeric):** Represents the torsion angle around the $N - C_\alpha$ bond.
- **bfactor (numeric):** Indicates the Debye-Waller factor.
- **areaSAS (numeric):** Reflects the solvent-accessible surface area.
- **areaSES (numeric):** Specifies the solvent-excluded surface area.
- **kdHydrophobicity (numeric):** Quantifies the hydrophobicity of the amino acid.
- **ConsurfDB (numeric):** Measures the degree of conservation of the amino acid.
- **degree (numeric):** Quantifies the number of connections of an amino acid within the protein network.
- **betweenness (numeric):** Represents the number of shortest paths between all pairs of amino acids that traverse through a specific amino acid.
- **closeness (numeric):** Indicates the average distance between a specific amino acid and all other amino acids within the protein network.
- **burts (numeric):** Measures the dependency of

an amino acid on others within the protein structure.

- **pr (numeric):** Quantifies the importance of an amino acid within the protein network.
- **auth (numeric):** Represents the relevance of an amino acid within the protein network.
- **kcore (numeric):** Denotes the sub graph where all amino acids are interconnected with at least k other amino acids.
- **CalculatedSeverity (categorical):** Specifies the severity of Hemophilia A based on the levels of FVIII: severe ($FVIII : C < 1$), moderate ($1 \geq FVIII : C \leq 5$), and mild ($FVIII : C > 5$).

Table I shows the class distribution of the data. One can observe that the dataset does not have a significant imbalance with respect to the classes. The least represented class has 71 instances, and the most represented class has 192 instances.

TABLE I
HEMOPHILIA A DATASET CLASS DISTRIBUTION.

Calculated Severity	Instances	Distribution
Mild	192	46.26%
Severe	152	36.63%
Moderate	71	17.11%

IV. FVIII PROTEIN CHARACTERISTICS BASED CLASSIFICATION OF HEMOPHILIA A SEVERITY

To assess the classification problem related to the severity of hemophilia A, a comparative analysis of several traditional machine learning models was conducted. These models, selected to cover a broad spectrum of classification techniques.

The baseline model chosen for this analysis was the Zero Rules (ZeroR) classifier, which serves as a fundamental reference point by always predicting based on the database distribution. Following this, more sophisticated models were implemented, including Gaussian Naive Bayes, which assumes that features follow a normal distribution and are conditionally independent given the class label.

The K Nearest Neighbors (KNN) algorithm was also utilized, which classifies instances based on the majority class among the k-nearest neighbors in the feature space. Decision Trees were employed to recursively split the data based on feature values to maximize information gain or other criteria, providing an intuitive, yet powerful classification method.

Ensemble methods were a significant part of this study, starting with AdaBoost, which combines multiple weak classifiers to form a strong classifier by focusing on hard-to-classify instances. Bagging, another ensemble method, was used to reduce variance by averaging the predictions from multiple bootstrapped datasets.

Random Forest, an extension of bagging with decision trees, was included due to its robustness and ability to handle high-dimensional data effectively. Gradient Boosting was examined for its capability to produce a predictive model in the form of an ensemble of weak prediction models, typically decision trees, by optimizing the loss function.

Additionally, two popular gradient boosting frameworks, LightGBM and XGBoost, were included for their efficiency and performance in handling large datasets. The Support Vector Machine (SVM) was implemented for its effectiveness in high-dimensional spaces and its use of various kernel functions to handle non-linear classification tasks.

Finally, the Multi Layer Perceptron (MLP), a type of feedforward artificial neural network, was used despite not being a deep learning model, to evaluate its capability in capturing complex patterns through its multiple layers and non-linear activation functions.

Each of these classifiers combined with the experimental methodology brings a unique approach and set of assumptions to the problem, providing a comprehensive evaluation framework for determining the most effective model for classifying the severity of hemophilia A.

All classifiers in this analysis use standardization because the data exhibits numerical differences. Standardization is crucial as it ensures that each feature contributes equally to the model's performance by rescaling the data to have a mean of zero and a standard deviation of one. This process helps to mitigate the impact of varying scales and units, thereby improving the classifier's accuracy

and convergence speed during training. Without standardization, features with larger numerical values could unduly influence the model, leading to biased results and reduced overall performance.

In this study, a variety of classifiers were evaluated for model development, each accompanied by a distinct hyperparameter search space. Table II delineates the hyperparameters investigated for each classifier. By encompassing a broad spectrum of values, these hyperparameters facilitate an exhaustive exploration of the model's parameter space. Such meticulous hyperparameter tuning is essential for optimizing the models' performance on the given dataset. Excluding Gaussian Naive Bayes because it does not have hyperparameters.

TABLE II
PARAMETER GRID FOR CLASSIFIERS

Classifier	**Parameter Grid**
DummyClassifier	`strategy`: {'stratified', 'most_frequent', 'prior', 'uniform'}
KNeighborsClassifier	`n_neighbors`: {3, 5, 7, 9, 11} `weights`: {'uniform', 'distance'} `metric`: {'euclidean', 'manhattan', 'minkowski'}
DecisionTreeClassifier	`criterion`: {'gini', 'entropy'} `max_depth`: {None, 10, 20, 30, 40, 50} `min_samples_split`: {2, 5, 10} `min_samples_leaf`: {1, 2, 4}
RandomForestClassifier	`n_estimators`: {100, 200, 300} `criterion`: {'gini', 'entropy'} `max_depth`: {None, 10, 20, 30, 40, 50} `min_samples_split`: {2, 5, 10} `min_samples_leaf`: {1, 2, 4} `bootstrap`: {True, False}
AdaBoostClassifier	`n_estimators`: {50, 100, 150} `learning_rate`: {0.01, 0.1, 1, 10}
BaggingClassifier	`n_estimators`: {10, 50, 100} `max_samples`: {0.5, 0.7, 1.0} `max_features`: {0.5, 0.7, 1.0}
GradientBoostingClassifier	`n_estimators`: {100, 200, 300} `learning_rate`: {0.01, 0.1, 0.2} `max_depth`: {3, 5, 7} `min_samples_split`: {2, 5, 10} `min_samples_leaf`: {1, 2, 4} `subsample`: {0.8, 0.9, 1.0}
XGBClassifier	`n_estimators`: {100, 200, 300} `learning_rate`: {0.01, 0.1, 0.2} `max_depth`: {3, 5, 7} `colsample_bytree`: {0.3, 0.7, 1.0} `subsample`: {0.8, 0.9, 1.0} `gamma`: {0, 0.1, 0.2}
LGBMClassifier	`n_estimators`: {100, 200, 300} `learning_rate`: {0.01, 0.1, 0.2} `num_leaves`: {31, 50, 70} `max_depth`: {-1, 10, 20, 30} `subsample`: {0.8, 0.9, 1.0} `colsample_bytree`: {0.8, 0.9, 1.0}
SVC	`C`: {0.1, 1, 10, 100} `kernel`: {'linear', 'rbf', 'poly', 'sigmoid'} `gamma`: {'scale', 'auto'}
MLPClassifier	`hidden_layer_sizes`: {(50,), (100,), (50, 50), (100, 50)} `activation`: {'tanh', 'relu'} `solver`: {'sgd', 'adam'} `alpha`: {0.0001, 0.001, 0.01} `learning_rate`: {'constant', 'adaptive'}

V. EXPERIMENTAL METHODOLOGY

A. Resampling Strategy

To mitigate the issue of overly optimistic performance assessment, our study adopts a rigorous resampling strategy known as repeated nested cross-validation. This method entails iteratively partitioning the dataset into distinct sets for training, validation, and testing across multiple rounds. Crucially, nested cross-validation incorporates an inner loop dedicated to hyperparameter optimization, thereby ensuring a robust evaluation of model performance. By segregating the tuning process within the inner loop, the influence of occasional testing datasets on model training is effectively mitigated.

Furthermore, the utilization of multiple rounds of cross-validation facilitates a more nuanced application of statistical hypothesis testing. This approach enhances the reliability and generalizability of the model evaluation process. Figure 2 illustrates a schematic depiction of the division employed in nested cross-validation, employing 4 folds in the outer loop and 2 folds in the inner loop for demonstration purposes.

Experiments were conducted using a ten-fold cross-validation approach in the outer loop, where in each iteration, nine folds were used for training and one fold for testing. Within the inner loop, these nine training folds underwent a grid search with four-fold cross-validation to determine the optimal hyperparameters for each classifier. The identified

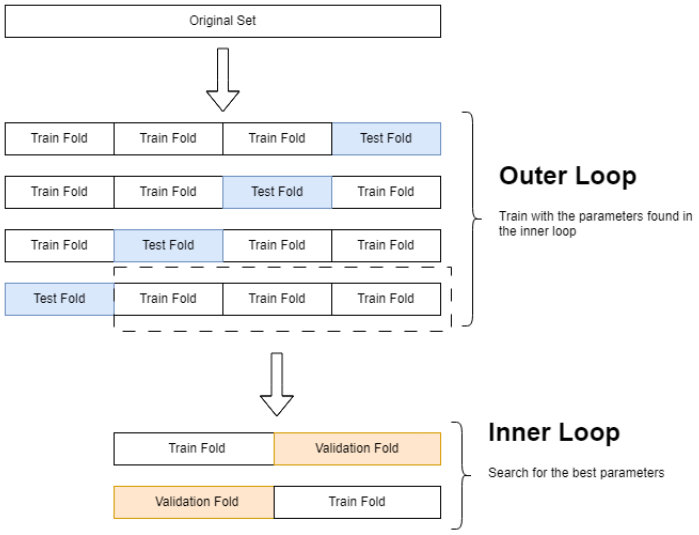

Fig. 2. Four Fold Nested Cross Validation Example.

best hyperparameters were then applied to the test fold to evaluate the classifier's performance. To mitigate the potential impact of an unfavorable sample division, this entire process was repeated three times, each time with a different partitioning of the data into folds.

B. Evaluation Metric and Statistical Tests

In datasets where class imbalances are minimal, accuracy emerges as the metric of choice for evaluating classifier performance. Its suitability lies in providing a comprehensive assessment across all classes, capturing the overall ability of classifiers to correctly classify instances. By prioritizing accuracy, the evaluation process aims to offer a holistic perspective on performance.

However, while accuracy is valuable in such scenarios, it may not always reveal the subtleties of classifier errors. For a more nuanced analysis, metrics like Macro One Versus Rest (OVR) Roc-Auc [2], Macro F1-Score and Macro recall [5]. These metrics unveil performance at the individual class level, offering a granular understanding of classifier behavior.

In comparing multiple classifiers, the conventional approach involves using hypothesis tests to determine statistical differences. Specifically, pairwise comparisons are extended when evaluating three or more methods. To rigorously assess performance disparities, the paired t-test and the Wilcoxon test are applied to the top three models based on accuracy.

The paired t-test assumes independence among samples, which is challenged in the resampling strategy utilized here due to overlapping training folds. To mitigate this, the corrected t-test proposed by Nadeau and Bengio [10] is employed.

Furthermore, the assumption of normal distribution in the corrected t-test necessitates a sufficiently large sample size. In this study, the resampling strategy encompasses three repetitions of 10 folds, yielding 30 results to meet the conditions for asymptotic normality, as outlined by the central limit theorem [3] .

In addition to the corrected t-test, the Wilcoxon test is employed to further validate comparisons, especially when dealing with data that may not conform to normal distribution assumptions.

TABLE III
RESULTS OF CLASSIFIER PERFORMANCE. BEST VALUES ARE HIGHLIGHTED IN BOLD.

Classifier	Accuracy	Accuracy(Std)	ROC AUC OVR	F1 Macro	Recall Macro
RandomForest	**0.641696**	0.074401	**0.756103**	0.540879	0.543627
GradientBoosting	0.623790	0.068989	0.735612	**0.551431**	0.548359
Bagging	0.622280	0.071305	0.725418	0.534433	0.537961
XGBoost	0.611963	0.080690	0.755442	0.540175	0.537943
LightGBM	0.604530	0.078620	0.742532	0.521414	0.523356
MLP	0.601413	0.063810	0.677539	0.494554	0.508636
DecisionTree	0.593419	0.067890	0.669237	0.544831	**0.552560**
SVM	0.577274	0.065609	0.644253	0.417032	0.459635
AdaBoost	0.557375	0.074609	0.641285	0.477084	0.480195
KNeighbors	0.486605	0.061810	0.567844	0.383259	0.397845
Dummy	0.462660	0.008282	0.500000	0.210861	0.333333
GaussianNB	0.461808	0.087069	0.613913	0.429252	0.443781

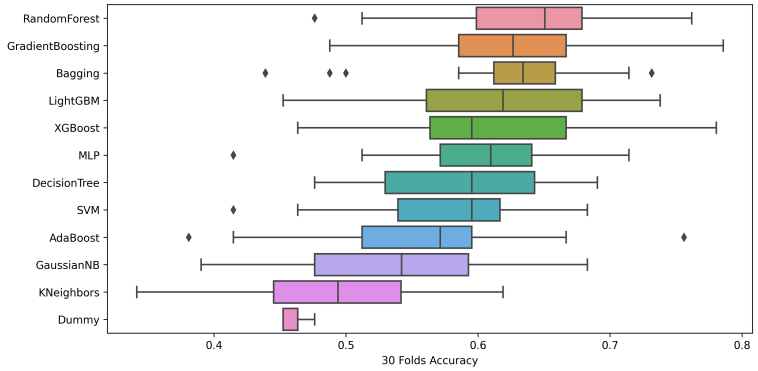

Fig. 3. Boxplots of the accuracies of the 12 models

TABLE IV
THE P-VALUES FROM BOTH THE PARAMETRIC CORRECTED PAIRED T-TEST AND THE NON-PARAMETRIC WILCOXON TEST ARE PROVIDED FOR THE TOP THREE MODELS. THE T-TEST RESULTS ARE DISPLAYED IN THE UPPER TRIANGULAR MATRIX, WHILE THE WILCOXON TEST RESULTS ARE PRESENTED IN THE LOWER TRIANGULAR MATRIX. A P-VALUE LESS THAN 0.05 INDICATES STATISTICALLY SIGNIFICANT DIFFERENCES AT A 95% CONFIDENCE LEVEL. SIGNIFICANT P-VALUES ARE HIGHLIGHTED IN BOLD FOR CLARITY AND EASE OF INTERPRETATION.

RandomForest	0.401	0.348
0.137	**GradientBoosting**	0.951
0.057	0.959	**Bagging**

VI. RESULTS AND ANALYSIS

This section presents the experimental results, comparing the metric values and analyzing the statistical differences between the outcomes to identify the best classification model.

Table III presents the performance metrics of various classifiers. The results indicate that the RandomForest classifier achieves the highest accuracy and ROC AUC OVR values, while the DecisionTree classifier has the best recall macro score. GradientBoosting leads in the F1 macro score. The Dummy classifier, used as a baseline, has the lowest performance across most metrics.

Figure 3 shows the boxplots of the accuracy's of the 12 models across 30 folds. The boxplots illustrate the distribution of accuracy scores for each classifier, with the median, interquartile range, and potential outliers clearly depicted.

It is evident that the RandomForest classifier

not only has a high median accuracy but also demonstrates relatively low variability, indicating consistent performance across different folds. GradientBoosting also shows strong performance with slightly higher variability. Bagging, LightGBM, and XGBoost classifiers exhibit comparable median accuracies but with varying degrees of spread, reflecting their stability in different cross-validation folds. On the lower end, classifiers like GaussianNB, KNeighbors, and the Dummy classifier show significantly lower accuracy and higher variability, underscoring their limited effectiveness for this dataset.

Table IV presents the p-values from both the parametric corrected paired t-test and the non-parametric Wilcoxon test for the top three models. The t-test results are displayed in the upper triangular matrix, while the Wilcoxon test results are presented in the lower triangular matrix. A p-value less than 0.05 indicates statistically significant differences at a 95% confidence level. Significant p-values are highlighted in bold for clarity and ease of interpretation.

The results indicate no statistically significant differences between the top three models (RandomForest, GradientBoosting, and Bagging) at the 95% confidence level. The lowest p-value observed is 0.057 (Wilcoxon test between RandomForest and Bagging), which is close to the threshold but not significant. This suggests that, based on the provided tests, there is no strong evidence to prefer one of these top-performing models over the others in terms of accuracy.

VII. Conclusion

This paper explored machine learning methodologies for classifying Hemophilia A severity on point mutation data in the FVIII protein. The classification models were thoroughly evaluated, with the Random Forest emerging as the most efficient model among those analyzed.

Although the Random Forest model yielded satisfactory results, it is important to note that its performance is not yet sufficient to replace the FVIII measurement test in blood for diagnostic purposes. Achieving this would require expanding the dataset with additional samples and possibly incorporating new features. Further investigation into advanced machine learning techniques is also necessary. This includes utilizing resampling methods to address data imbalance, exploring automated feature selection processes, and considering alternative classification models beyond those evaluated in this study. In conclusion, this research marks a significant step forward in developing classification models for hemophilia A severity.

Nonetheless, there remains substantial potential for enhancements and further studies to achieve even more precise and dependable outcomes in the future.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve language and readability. After using this tools, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

[1] A. E. Bowyer and R. C. Gosselin. Factor viii and factor ix activity measurements for hemophilia diagnosis and related treatments. *Seminars in Thrombosis and Hemostasis*, 2022. doi: 10.1055/s-0042-1758870.

[2] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(96)00142-2.

[3] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, Los Angeles, 2002.

[4] Giancarlo Castaman and Davide Matino. Hemophilia a and b: molecular and clinical similarities and differences. *Haematologica*, 104(9):1702–1709, Aug. 2019. doi: 10.3324/haematol.2019.221093.

[5] Nancy Chinchor. MUC-4 evaluation metrics. In *Fourth Message Understanding Conference (MUC-4): Proceedings*

of a Conference Held in McLean, Virginia, June 16-18, 1992, 1992. URL https://aclanthology.org/M92-1002.

[6] A. A. Ferreira, I. V. Brum, J. V. L. Souza, and I. C. G. Leite. Cost analysis of hemophilia treatment in a brazilian public blood center. *Cadernos de Saúde Coletiva*, 2020. doi: https://doi.org/10.1590/1414-462X202028040484. Ahead of Print.

[7] World Federation Of Hemophilia. Wfh, 2021. report on the annual global survey, 2021. URL http://www.wfh.org.

[8] Thiago J. S. Lopes, Renata Rios, Tiago Nogueira, and Rodrigo F. Mello. Prediction of hemophilia A severity using a small-input machine-learning framework. *NPJ Systems Biology and Applications*, 7(1):22, May 2021. doi: 10.1038/s41540-021-00183-9.

[9] Agnieszka Mazurkiewicz-Pisarek, Grażyna Płucienniczak, Tomasz Ciach, and Andrzej Płucienniczak. The factor viii protein and its function. *Acta Biochim Pol*, 63(1):11–16, 2016. doi: 10.18388/abp.2015$_1$056.

[10] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Advances in neural information processing systems*, 12, 1999.

[11] Jacky Chi Ki Ngo, Mingdong Huang, David A. Roth, Barbara C. Furie, and Bruce Furie. Crystal structure of human factor viii: Implications for the formation of the factor ixa-factor viiia complex. *Structure*, 16(4):597–606, 2008. ISSN 0969-2126. doi: https://doi.org/10.1016/j.str.2008.03.001.

[12] Alok Srivastava, Elena Santagostino, Alison Dougall, Steve Kitchen, Megan Sutherland, Steven W. Pipe, Manuel Carcao, Johnny Mahlangu, Margaret V. Ragni, Jerzy Windyga, Adolfo Llinás, Nicholas J. Goddard, Richa Mohan, Pradeep M. Poonnoose, Brian M. Feldman, Sandra Zelman Lewis, H. Marijke van den Berg, Glenn F. Pierce, and the WFH Guidelines for the Management of Hemophilia panelists and co-authors . Wfh guidelines for the management of hemophilia, 3rd edition. *Haemophilia*, 26(S6):1–158, 2020. doi: https://doi.org/10.1111/hae.14046.

[13] James S Stonebraker, Paula H B Bolton-Maggs, J Michael Soucie, Paul Walker, and Charles Brooker. A study of variations in the reported hemophilia a prevalence around the world. *Haemophilia*, 16(1):20–32, 2010.

[14] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.