Dear Editor,

Digital Health Journal,

Thank you for taking the time to review our paper. We appreciate the valuable feedback you provided, which we believe will greatly improve the quality of our work.

We are pleased to hear that you found our research topic interesting and relevant and that the structure and organization of our paper were clear and logical. Your constructive criticism on the other hand has been very helpful in identifying areas where we can make improvements.

We have taken your comments into consideration and have made the following changes to our paper:

We have revised our introduction to provide a more comprehensive overview of the problem and the motivation behind our study.

We have clarified the methodology section by providing more detailed descriptions of the statistical analyses used in our study.

We have added more references to support our arguments and to provide a more complete picture of the current state of research in our field.

We have made some changes to the language and phrasing throughout the paper to improve clarity and readability.

We hope that these changes address your concerns and that you find our revised paper to be of high quality. We appreciate your time and effort in reviewing our work and providing us with such thoughtful feedback.

Thank you again for your assistance in improving our manuscript.

Sincerely,

Dr Fahad Alturise

***Reviewer: 1***

***1.*** ***What is the innovation of this work?***

***Answer***:

The importance of the work is realized as an in-silico model in comparison with In-vitro and in-vivo analysis which are typically both time-consuming and expensive, due to the need for laboratory facilities, specialized equipment, and trained personnel. In addition, these types of experiments can also be ethically and legally complex, as they often involve testing on live animals or human subjects. In-silico analysis, on the other hand, uses computer simulations and modeling to study biological processes, making it a more cost-effective and efficient alternative. It also eliminates the need for animal testing and can provide a more controlled and reproducible environment for testing. As a result, in-silico analysis has become a popular tool for researchers to study various biological processes and to help guide their experimental design. Furthermore, the study aimed to improve upon the existing methods yielding higher accuracy. Firstly, as per the observed evaluation of Hemolytic-Pred in comparison with existing models, we infer that Hemolytic-Pred gave the highest scores for ACC, Sp, Sn, MCC, and AUC-ROC. This demonstrates that the proposed method can be considered the most accurate and reliable predictor till now.

***2.*** ***The Introduction needs to be reorganized. It is not necessary to describe the background of hemolytic proteins so much. Authors need to introduce more related works and their limitations.***

***Answer: The introduction has been extensively revised. More related work has been incorporated as per the advice of the esteemed reviewer.***

***3.*** ***For each figure, the authors should give more details in the title.***

***Answer:***

***The whole manuscript has been rigorously checked such that it contains details regarding each figure.***

***4.*** ***For figure 3, why do the authors give the illustration of XGBoost and not for other classifiers? If they give the illustration of XGBoost, they also need to give those of other classifiers.***

***Answer:***

***We are very grateful to the esteemed reviewer for his valuable comments. More illustrations for XGBoost, Ada boost, and Neural Network have been incorporated into the manuscript.***

***5.*** ***In results and discussions, more insightful discussions should be given. For example, why does the XGBoost perform better than other classifiers and existing methods?***

***XG Boost is an iterative learning method where the model self-analyzes its mistakes and gives more weightage to misclassified data points in the next iteration, making it a reliable model. XG Boost uses a similarity score to prune trees and prevent overfitting. It is a better option for unbalanced datasets, such as in fraud detection, and is more efficient in optimizing hyperparameters compared to Random Forest. In contrast, Random Forest may give preference to classes with more participation, making XG Boost a more preferable choice in situations like Poisson regression and rank regression.***

***6.*** ***The authors used five feature descriptors to represent hemolytic proteins. So, which features provide more contributions to the final prediction?***

***Answer:***

***All the computed features in previous phases are merged into a feature vector, and for each data sample. It formulated a feature input matrix of fixed size for any arbitrary sequence irrespective of its length. The matrix formed by combining the feature vector obtained for each sample is further input to the machine learning classifiers for training and further evaluation.***

***All the computed features from the previous phases are merged into a feature vector. For each data sample, a feature input matrix of fixed size is formulated, regardless of the sequence's length. The matrix, which is formed by combining the feature vector obtained for each sample, is further input to the machine learning classifiers for training and evaluation***

***7.*** ***The Webserver is a plus for this job. But there is something wrong with the site, it doesn't open, and the author needs to check the back-end code properly.***

***Answer: The Webserver has been thoroughly checked and fixed. It is now publicly available at the given link.***

***8.*** ***Here authors developed the model on the imbalanced dataset. I suggest authors to show the model performance on a balanced dataset and compare the performance with other methods as well.***

***Answer:***

***We are very grateful to the reviewer for providing such valuable comments. We have further extend the experimentation to include balanced dataset for testing and validation.***

***9.***

***Authors should also correlate their study with (i) 10.1093/bib/bbab041 and (ii) 10.1093/bioinformatics/btac006 and cite them appropriately.***

***Answer: Done 21,71 is cited.***

***10.***  ***I would suggest authors add a section highlighting the limitations of the current study. Answer:***

***Limitation of the work has been mentioned in the Conclusion section***

***Reviewer: 2***

***1.*** ***The introduction should be refined as it has inconsistencies in information. Also, the recent advances in bioinformatics should be discussed.***

***Answer: We are very grateful to the reviewer for his valuable comments. We have refined and revised the introduction extensively to address all the issues raised by the esteemed reviewer.***

***2.***

***The language of the article needs to be reviewed.***

***Answer: The Manuscript has been thoroughly revised for Language errors.***

**3.** **The authors should move the “Formulation of Metrics” section to methods.**

***Answer: The change has been incorporated into the manuscript.***

***4.*** ***Graphical illustration of other classifiers should be provided, as it is provided for XGBoost.***

***Answer: illustration for XGBoost, Adaboost, and Neural Network has been incorporated***

***5.*** ***All equations and their symbols should be explained.***

***Answer:***

***6.*** ***Limitations of the study should be discussed as well in the manuscript.***

***Answer: Limitation of the work has been mentioned in Conclusion section***

***Reviewer: 3***

**Please ensure that you provide an Abstract under following headings: Objective, Methods, Results and Conclusion.**

Objective:

The objective of this study is to propose a novel method called Hemolytic-Pred for identifying hemolytic proteins based on their sequences, using statistical moment-based features, along with position-relative and frequency-relative information.

Methods:

The authors used various machine learning algorithms for classification purposes, and four different validation approaches were opted for the exhaustive evaluation of the model. The Hemolytic-Pred webserver is available for further analysis.The Hemolytic-Pred webserver is available at

\href{http://ec2-54-160-229-10.compute-1.amazonaws.com/}{\color{blue}{http://ec2-54-160-229-10.compute-1.amazonaws.com/}}

Results:

XGBoost outperformed the other six classifiers with a value of 0.99, 0.98, 0.97, and 0.98 for self-consistency test, 10-fold cross-validation, Jackknife test, and independent set test, respectively. The proposed method with the XGBoost classifier is a workable and robust solution for predicting hemolytic proteins efficiently and accurately.

Conclusion:

The proposed method of Hemolytic-Pred with XGBoost classifier is a reliable tool for the timely identification of hemolytic cells and diagnosis of various related severe disorders. The application of Hemolytic-Pred can yield profound benefits in the medical field.

Please ensure that you discuss in full the limitations of this study in the Discussion

***Reviewer: 1***

***1.*** ***What is the innovation of this work?***

***Answer***: The main aim of our study was to improve upon the existing methods with higher accuracy. For Hemolytic-Pred, the evaluation has already been evaluated, thus, to compare the results with previous methods, In comparison, we observe that Hemolytic-Pred gave the highest scores for ACC, Sp, Sn, MCC, and AUC-ROC, as compared to the existing methods. This infers that the proposed method can be considered the most accurate and reliable predictor till now. *<<<Include some stuff regarding the significance of Hemolytic proteins in terms of any disease-related research. Also mention in-vitro and in-vivo analysis is expensive and time taking that's why researchers often opt for in-silico solutions for protein characterization>>>*

The main aim of our study was to improve upon the existing methods with higher accuracy. Firstly, as per **the** observed evaluation of Hemolytic-Pred in comparison with existing models, we infer that Hemolytic-Pred gave the highest scores for ACC, Sp, Sn, MCC, and AUC-ROC. This **demonstrates** that the proposed method can be considered the most accurate and reliable predictor till now. **Furthermore, its importance is realized as an in-silico model in comparison with In-vitro and in-vivo analysis which are typically both time-consuming and expensive, due to the need for laboratory facilities, specialized equipment, and trained personnel. In addition, these types of experiments can also be ethically and legally complex, as they often involve testing on live animals or human subjects. In-silico analysis, on the other hand, uses computer simulations and modeling to study biological processes, making it a more cost-effective and efficient alternative. It also eliminates the need for animal testing and can provide a more controlled and reproducible environment for testing. As a result, in-silico analysis has become a popular tool for researchers to study various biological processes and to help guide their experimental design.**

***2.*** ***The Introduction needs to be reorganized. It is not necessary to describe the background of hemolytic proteins so much. Authors need to introduce more related works and their limitations.***

***Answer: <<< Have u made changes to the introduction?? Mention here the changes u made >>>***

***No changes made by me.***

***3.*** ***For each figure, the authors should give more details in the title.***

***Answer: <<< Each figure shall be described and cited within the text. You must explain what each figure illustrates>>>***

***DONE IN THE MANUSCRIPT.***

***4.*** ***For figure 3, why do the authors give the illustration of XGBoost and not for other classifiers? If they give the illustration of XGBoost, they also need to give those of other classifiers.***

***Answer: illustration For XGBoost, Ada boost and Neural Network has been incorporated into the manuscript.***

***5.*** ***In results and discussions, more insightful discussions should be given. For example, why does the XGBoost perform better than other classifiers and existing methods?***

Boosting happens to be iterative learning which means the model will predict something initially and self analyze its mistakes as a predictive toiler and give more weightage to the data points in which it made a wrong prediction in the next iteration. After the second iteration, it again self-analyses its wrong predictions and gives more weightage to the data points which are predicted as wrong in the next iteration. This process continues as a cycle. Hence technically, if a prediction has been made, there is at most surety that it did not happen as a random chance but with a thorough understanding and patterns in the data. Such a model that prevents the occurrences of predictions with a random chance is trustable most of the time.

· **XG Boost straight away prunes the tree with a score called “Similarity score” before entering into the actual modeling purposes.** It considers the “Gain” of a node as the difference between the similarity score of the node and the similarity score of the children. If the gain from a node is found to be minimal then it just stops constructing the tree to a greater depth which can overcome the challenge of overfitting to a great extend. Meanwhile, the Random forest might probably overfit the data if the majority of the trees in the forest are provided with similar samples. If the trees are completely grown ones then the model will collapse once the test data is introduced. Therefore major consideration should be given to distributing all the elementary units of the sample with approximately equal participation to all trees.

· **XG Boost is a good option for unbalanced datasets but we cannot trust random forest in these types of cases.** In applications like forgery or fraud detection, the classes will be almost certainly imbalanced where the number of authentic transactions will be huge when compared with unauthentic transactions. In XG Boost, when the model fails to predict the anomaly for the first time, it gives more preferences and weightage to it in the upcoming iterations thereby increasing its ability to predict the class with low participation but we cannot assure that random forest will treat the class imbalance with a proper process.

· **One of the most important differences between XG Boost and Random forest is that the XG boost always gives more importance to functional space when reducing the cost of a model while Random Forest tries to give more preferences to hyperparameters to optimize the model.** A small change in the hyperparameter will affect almost all trees in the forest which can alter the prediction. Also, this is not a good approach when we expect test data with so many variations in real-time with a pre-defined mindset of hyperparameters for the whole forest but XG boost hyperparameters are applied to only one tree at the beginning which is expected to adjust itself in an efficient manner when iterations progress. Also, the XG boost needs only a very low number of initial hyperparameters (shrinkage parameter, depth of the tree, number of trees) when compared with the Random forest.

· When the **model is encountered with a categorical variable** with a different number of classes then there lies a possibility that a Random forest may give more preferences to the class with more participation. You can relate this point to point 3.

· XG Boost may more preferable in situations like Poisson regression, rank regression, etc. This is because trees are derived by optimizing an objective function.

***Summarized answer***

***XG Boost is an iterative learning method where the model self-analyzes its mistakes and gives more weightage to misclassified data points in the next iteration, making it a reliable model. XG Boost uses a similarity score to prune trees and prevent overfitting. It is a better option for unbalanced datasets, such as in fraud detection, and is more efficient in optimizing hyperparameters compared to Random Forest. In contrast, Random Forest may give preference to classes with more participation, making XG Boost a more preferable choice in situations like Poisson regression and rank regression.***

***6.*** ***The authors used five feature descriptors to represent hemolytic proteins. So, which features provide more contributions to the final prediction?***

***Answer: <<<Here explain not just one type of feature is used as input for each classifier. Rather, all the features are combined to form a larger feature vector. |Each sequence no matter how large or small is transformed into a fixed-size feature vector most suitable for training of any machine learning model.>>>***

***7.*** ***The Webserver is a plus for this job. But there is something wrong with the site, it doesn't open, and the author needs to check the back-end code properly.***

***Answer: checked and fixed.***

***8.*** ***Here authors developed the model on the imbalanced dataset. I suggest authors to show the model performance on a balanced dataset and compare the performance with other methods as well.***

***Answer: Done for balance data too and update in overleaf too.***

***9.***

***Authors should also correlate their study with (i) 10.1093/bib/bbab041 and (ii) 10.1093/bioinformatics/btac006 and cite them appropriately.***

***Answer: I am unable to download the paper. <<< Do not download them. Cite them appropriately within text based on its context. THIS IS VERY VERY IMPORTANT>>>>>***

***DONE 19,69 IS CITED IN THE MANUSCRIPT.***

***10.***  ***I would suggest authors add a section highlighting the limitations of the current study. Also, I would be interested in knowing how this approach could potentially take the findings of the study to clinical applications.***

***Answer:*** The biological sequence data increases day by day at

high speed in a different type of database like the Swiss Prot database, in the future, the space to improve efficiency in this field still exists due to the increasing number of datasets in available databases. *<<< Explain the diseases to which hemolytic proteins are linked using some reference. Then explain that scientists working on the treatment or cure of this disease would require to characterize a protein as hemolytic or non-hemolytic protein. This is where our accurate and robust model would help the researcher. This model can help in drug design and discovery for the above-mentioned diseases>>>*

***Red part of the question please sir explain.***

***A number of diseases are associated with hemolytic proteins, including sickle cell anemia, G6PD deficiency, Hemolytic Uremic Syndrome, Thalassemia, Autoimmune Hemolytic Anemia (AIHA), Pyruvate Kinase Deficiency, Spherocytosis, Glucose-6-Phosphate Dehydrogenase (G6PD) deficiency, and Paroxysmal Nocturnal Hemoglobinuria (PNH). In order to treat or cure the aforementioned diseases, scientists would need to determine whether a protein is hemolytic or non-hemolytic. The researcher could benefit from our reliable and accurate model in this case. Using this model, it might be possible to design and discover drugs for the diseases listed above.***

***Reviewer: 2***

***1.*** ***The introduction should be refined as it has inconsistencies in information. Also, the recent advances in bioinformatics should be discussed.***

***Answer: <<< Have u revised the intro likewise>>>***

***2.***

***The language of the article needs to be reviewed.***

***Answer: Reviewed via using Grammarly.***

**3.** **The authors should move the “Formulation of Metrics” section to methods.**

***Answer: <<< DO it>>>***

***DONE PLEASE CHECK IN THE MENUSCRIPT.***

***4.*** ***Graphical illustration of other classifiers should be provided, as it is provided for XGBoost.***

***Answer: illustration for XGBoost, Adaboost, and Neural Network has been incorporated***

***Answer: done for 3.***

***5.*** ***All equations and their symbols should be explained.***

***Answer:***

***6.*** ***Limitations of the study should be discussed as well in the manuscript.***

***Answer:*** The biological sequence data increases day by day at

high speed in a different type of database like the Swiss Prot database, in the future, the space to improve efficiency in this field still exists due to the increasing number of datasets in available databases.

**DATASET:**

The hemolytic protein dataset was collected from the UniProtKB-SwissProt using the keyword “Hemolysis [KW-0354]”.

The protein sequences were collected on February 2, 2021. The correctness of data is the key attribute that drives

the performance of any desired predictor. A specific well-defined set of rules are used to collect robust and accurate dataset as described in previous many studies. Based on these criteria data set is collected that is best in quality, informative, accurate, and diverse. In the first stage, the sequences with ambiguous annotations like “fragment”, “potential”, “probable”, “probably”, “maybe”, or “by similarity” were excluded from the data set.

We made two datasets one is ImBD (Imbalance Data set ) and BD(balance Data set).

**ImBD (Imbalance Data set ):**

This yielded a total of 7,107 Hemolytic proteins out of which, only 946 were reviewed, thus, only these reviewed sequences were included in the positive dataset. Similarly, the “non-Hemolysis” proteins were also collected from UniProtKB SwissProt, by using a converse query, and a total of 980 “non-Hemolysis” reviewed proteins were collected to be used as a negative dataset.

After retrieving data from UniProt, redundancy from the dataset was reduced by using CD-HIT program with a threshold value of 0.7, i.e., sequences having similarity of more than 70% were excluded from dataset (19). This yielded 329 clusters of positive dataset, and 891 clusters of negative dataset. Thus, only 1 representative from each cluster was used,and the unbalanced dataset comprised 329 positive sequences and 891 negative sequences.

**BD (balance Data set ):**

In the balance data set, we get positive and negative Hemolytic proteins collected from UniProtKB SwissProt, which we collected to be used as a negative dataset comprising 329 positive sequences and 330 negative sequences.