**LETTER OF TRANSMITTAL**

June, 2022

To

Supriya Sarker

Lecturer

Department of Computer Science & Engineering

World University of Bangladesh

Plot: 5-8, Avenue 6, Lake Drive Road, Sector 17/H, Uttara

Dhaka 1230, Bangladesh.

**Subject: Submission of Thesis Dissertation**

Dear Madam,

We are pleased to submit the dissertation entitled **“Explaniable Diabetes Prediction”**. It was a great pleasure to work on such an important topic. The dissertation is prepared according to the requirements and guidelines of the Department of Computer Science & Engineering, World University of Bangladesh.

We believe that the dissertation will help you in evaluating our thesis work. It would be a great pleasure for us to interpret any part or the entire project when necessary.

Sincerely yours Sincerely yours

\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Md. Masum Billah Opu Md. Saiham

WUB 03/18/42/2671 WUB 03/18/42/2674



**World University of Bangladesh**

**DECLARATION**

We hereby, solemnly declare that the thesis work entitled **“Explaniable Diabetes Prediction”,** has been supervised by Supriya Sarker, Lecturer of the Department of Computer Science & Engineering, World University of Bangladesh. We ensure that the project report has not been submitted either in whole or part for any degree or Diploma in any university previously.

We hereby, warrant that the work we have presented does not breach any existing copyright rule. We further undertake to indentify the university against any loss or damage arising from breach of the foregoing obligation.

Sincerely yours Sincerely yours

\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Md. Masum Billah Opu Md. Saiham

WUB 03/18/42/2671 WUB 03/18/42/2674



**Department of Computer Science & Engineering**

**World University of Bangladesh**

**CERTIFICATE**

I hereby, certify that the Thesis Dissertation on **“Explaniable Diabetes Prediction Model”,** is a confide report of thesis done by **Md. Masum Billah Opu and Md. Saiham** for partial fulfillment of the requirements for award of the degree of the Bachelor of Science in Computer Science and Engineering from World University of Bangladesh.

The thesis report has been carried out under my guidance and is a record of the bona-fide work carried out successfully by the students.

Supervisor

\_\_\_\_\_\_\_\_\_\_\_\_

Supriya Sarker

Lecturer

Department of Computer Science & Engineering

World University of Bangladesh (WUB)

**ACKNOWLEDGEMENTS**

We are extremely grateful and remain indebted to Almighty ALLAH who has guided us in ventures to complete this project. We are thankful for the grace and the help received from Him. It is our pleasure to thank our honorable **Vice Chancellor of WUB, Prof. Dr. Abdul Mannan Choudhury,** to whom we owe a lot for giving us an opportunity to complete our project.

We are especially thankful to our project supervisor, Supriya Sarker, Lecturer of the Department of Computer Science & Engineering, the World University of Bangladesh, for helping us to prepare a report on **“Explaniable Diabetes Prediction”,** and her continuous guidance. Her instructions and guidance have been of extreme help to us. We have consulted her and she always answered with utmost patience and perseverance.

We are greatly indebted to all of our teachers and staffs of the Department of Computer Science & Engineering and other departments of the World University of Bangladesh for their kind assistance in the accomplishment of this thesis work.

We are thankful to all of our friends, who are indeed, very close to our heart. We will always be indebted for the love and care that our parents have shown on us. We are very lucky to get such caring and loving parents. We feel the same way for our parents who have been the source of our inspiration.

Sincerely yours Sincerely yours

\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Md. Masum Billah Opu Md. Saiham

WUB 03/18/42/2671 WUB 03/18/42/2674

**ABSTRACT**

Diabetes is becoming a prolific disease and a global pandemic. The capability to detect diabetes early can improve the scope of disease management. For disease prognosis, artificial intelligence techniques are used widely. In this paper, probability and prediction for diabetes are discussed along with its model explanability and interpretability. We have used two different datasets which have been collected from UCI repository. For model interpretation we have used three XAI tools which are- Shapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and Explain Like I’m 5 (ELI5).

In the prior research study, many researchers have proposed several predictive diabetic models. But the models they have proposed are black box type and not interpretable. A robust interpretable model representing a good correlation with cardinal features of diabetes can provide more confidence to medical practitioners and patients in the model prognostics.

We have emphasised on the explainability an interpretability of our predictive model. First, we have processed our data and split the data into two sets, train data and test data. Then we have built our trained model by feeding the train data on the machine learning models. For this work we have used five machine learning algorithms which are- Logistic Regression, Random Forest, XgBoost Classifier, Decision Tree, Support Vector Machine.

Then we have evaluated our built model using the test data. The evaluation metrices we have used to compare different machine learning algorithms are precision, recall, f-1 score and accuracy. We have used three XAI tools to enhance the explainability of our model.

We have proposed an explainable diabetes prediction model. Our model can predict successfully whether a patient is in any risk of having diabetes or not. We have also explained the reason behind our model’s prediction using various XAI tools. We plan to use this model in future for prediction of other prolonged chronic diseases such as cancer, heart disease and arthirtis etc.

**Table of Contents Page No**

Letter of Transmittal …………………………………………………………………….... I

Declaration ……………………………………………………………………………...... II

Certificate ……………………………………………………………………………….... III

Acknowledgements ……………………………………………………………….............. IV

Abstract …………………………………………………………………………................. V

Table of Contents ……………………………………………………………………......... VI

List of Figures …………………………………………………………………………… VIII

**Chapter 1: Introduction………………………………………………………………… 1-2**

* 1. Introduction……………………………………………………………………........... 1
  2. Objectives………………………………………………………………….................. 2
  3. Justification of Study……………………………………………………………….... 2
  4. Scopes of Study…………………………………………………………………......... 2

**Chapter 2: Literature Review…………………………………………………………… 3-5**

* 1. Background Study ………………………………………………………….…........... 3
  2. Related Works ………...…………………………………………………………… 4-5

**Chapter 3: Methodology………………………………………………………………. 6-8**

* 1. Methodology …………...……………………………………………………............ 6
  2. Description of Methodology …………...……………………………………………. 7
     1. Data Preprocessing……………...…………………………………………………. 7
     2. Missing Values Removal…………………………………………………………… 8
     3. Splitting of Data……………………………………………………………………. 8
     4. Apply Machine Learning Algorithms………………………………………………. 8
     5. Explain with AI tools………………………………………………………………… 8

**Chapter 4: Requrement Analysis & Model Description……………………………. 9-13**

* 1. Dataset………………………………………………………………..................... 9-10

4.1.1 Dataset I - Pima Indian Dataset ……………………………………………………… 9

4.1.2 Dataset II – Diabetes Dataset of Sylhet…………………………………………… 9-10

* 1. Language and Tool Requriement………………………………………………… 10

4.2.1 Python………………………………………………………………………………. 10

4.2.2 Jupyter Notebook…………………………………………………………………… 10

* 1. Machine Learning Algorithms…………………………………………………….. 11

4.3.1 Logistic Regression.……………………………………………………….................. 11

4.3.2 Random Forest……………………………………………………………………….. 11

4.3.3 Decision Tree……………………………………………………………………….... 11

4.3.4 XgBoost Classifier………………………………………………………………….... 11

4.3.5 Support Vector Machine……………………………………………………………… 12

4.4 Explainable AI Tools…………………………………………………………………… 12

4.5 Description of the Proposed Model…………………………………………………….. 12

**Chapter 5: Result and Analysis………………………………………………………..14-19**

* 1. Confusion Matrix………………………………………………………...................14-15
  2. Evaluation Metrices…………………………………..…………………………….15-16
  3. Interpretation with XAI Tools……………………………………………………...17-19
     1. SHAP…………………………………………………………………………………17
     2. LIME…………………………………………………………………………………18
     3. ELI5…………………………………………………………………………………..19

**Chapter 6: Conclusion…………………………………………………………………… 20**

* 1. Conclusion …………………………………………………………………………….20
  2. Limitations …………………………………………………………………………….20
  3. Future Works…………………………………………………………………………..20

**References……………………………………………………………………………. 21-22**

**List of Figures**

|  |  |  |
| --- | --- | --- |
| Number of Figure | Name of Figure | Pages |
| 3.1 | Methodology of Machine Learning Model | 7 |
| 5.1 | Confusion Matrix of ML algorithms for Dataset-I | 15 |
| 5.2 | Confusion Matrix of ML algorithms for Dataset-II | 15 |
| 5.3 | SHAP plot- patient with high-risk score | 17 |
| 5.4 | LIME plot- Explaining feature contribution to patients with high and low risk probability | 18 |
| 5.5 | ELI5 – Feature Importance Plot | 19 |