LGURJCSIT
Volume 06, Issue 03, Article ID 327, 11 pages
https://doi.org/10.54692/lgurjcsit.2022.06003327

# An Intelligent Healthcare System for Detecting Diabetes using Machine Learning Algorithms

Hassan Kaleem[1], Saman Liaqat[1], Malik Tahir Hassan[2], Aneela Mehmood[2], and Allah Dita[3]
[1, 2, 3]School of Systems and Technologies, University of Management and Technology Lahore, Pakistan
[2]Computer Science Department Lahore Garrison University, DHA Phase 6, Pakistan
Hassanrao.hr@gmail.com, Samanlibran@gmail.com

(Received 04 July 2022; Accepted 17 July 2022; Published 25 July 2022)

***ABSTRACT-Human disease prediction is specifically a struggling piece of work for accurate and on-time treatment. Around the world, diabetes is a hazardous disease. It affects the various essential organs of the human body, for example, nerves, retinas, and eventually the heart. By using models of machine learning algorithms, we can recommend and predict diabetes on various healthcare datasets more accurately with the assistance of an intelligent healthcare recommendation system. Not long ago, for the prediction of diabetes, numerous models and methods of machine learning have been introduced. But despite that, enormous multi-featured healthcare datasets cannot be handled by those systems appropriately. By using Machine Learning, an intelligent healthcare recommendation system is introduced for the prediction of diabetes. Ultimately, the model of machine learning is trained to predict this disease along with K-Fold Cross validation testing. The evaluation of this intelligent and smart recommendation system is depending on datasets on diabetes and its execution is differentiated from the latest development of previous literature. Our system accomplished 99.0% of efficiency with the shortest time of 12 Milliseconds, which is highly analyzed by the previously existing models of machine learning. Consequently, this recommendation system is superior for the prediction of diabetes to the previous ones. This system enhances the performance of automatic diagnosis of this disease. The code is available at (https://github.com/RaoHassanKaleem/Diebetes-Detection-using-Machine-Learning-Algorithms).***

**Keywords:** Intelligent healthcare System, Diabetes Detection, Machine Learning Algorithms, K-Fold Cross-validation testing.

## INTRODUCTION

Diabetes is a hazardous disease through which almost 9% of the world's population suffered in the past, 422 million people are still fighting this disease worldwide, [1] 1.5 million people are dead already, and 3.7 million people are in quest of death from diabetes and high blood pressure (BP). This life-threatening disease affects the essential parts of the human body like the eyes, kidneys, lungs and heart. Nowadays youth is getting highly affected by diabetes as reported in [2]. Since diabetes has a massive effect on global health and the economy, it is essential to enhance methods for predicting and preventing diabetes [3].

Before this medical practitioners use to predict this disease manually and by using automatic devices. These measurements for predicting diabetes have some pros and cons as well. Nobody can predict this disease earlier, even if it was an experienced medical practitioner, because of the hideous side effects occurred in the human body [4]. But with the assistance of an intelligent healthcare system along with machine learning algorithms, we can predict diabetes early with minimum errors. In a recent development of biotechnology, high computing is steadily taking part in efficient and reliable e-healthcare disease prediction and data collection. The accurate model building would lead to a reliable and efficient system

by collecting most suitable data electronically [5]. In this paper, we have used 40 algorithms of machine learning to detect diabetes. After conducting our experiment, we have found that LGBM classifier has taken the shortest time (12 milliseconds) and provided us with the most accurate results as compared to other algorithms used in this paper.

Moreover, the K-Fold algorithm is used to measure cross-validation testing scores. After that means and the standard deviation are calculated to measure the actual accuracy of the models.

In recent research, various recommendation systems are already introduced for healthcare and disease prediction. For the prediction of diabetes, our essential contribution to this research is to enhance the accuracy and efficiency of the system. We have gathered data electronically from two different sources PIMA Indian diabetes dataset [6] and the Frankfurt Germany hospital's dataset [7]. Eventually, we proposed a better healthcare system for accurate and efficient prediction and detection of diabetic patients. The distribution of a paper is given below: Section 2 elaborates on the latest research conducted previously from the literature. Section 3 is the materials and methods used in this research. Section 4 provides training and testing of the dataset and the results it generated. Section 5 presents a comparative analysis of the latest research on diabetes detection using machine learning algorithms. Section 6 presents a dataset and code that is available on GitHub [8]. Section 7 describes

the conclusion and future research direction, and the final section leads to references.

## LITERATURE REVIEW

For an accurate and on-time treatment, a human disease prediction is specifically a struggling piece of task. Diabetes is a hazardous disease, around the world. It affects the various essential organs of the human body, for example, nerves, retinas, and eventually the heart. By using algorithms of machine learning, we can detect and predict diabetes on various healthcare datasets more accurately with the assistance of an intelligent healthcare system.

For the prediction of diabetes, numerous models and methods of machine learning have been introduced in previous research. Despite that, enormous multi-featured healthcare datasets cannot be handled by them appropriately. Various researchers have a great amount of contribution to detecting diabetes disease. They have used techniques related to artificial intelligence and machine learning for predicting diabetes disease. We can gather data effortlessly with the assistance of artificial intelligence. We can now easily detect and predict diabetes after gathering this big data from datasets available online. In previous diabetes prediction, a model of the k-nearest neighbor classifier was used, and 85.6% of novelty was assured with the comparison of the support vector machine [9]. In another paper, the convolution neural network technique with the comparison of the model of linear regression and multilayered perceptron was used for detecting diabetes [10]. Early analysis of detecting diabetes by using feature selection technique along with SVM classifier, the results carried out were compared with random forest, Naive Bayes, Decision Tree and KNN models, 77.73% of accuracy was achieved [11]. For comparing the results, various models of machine learning were used to authenticate them.

Bloodless techniques along with computational tools were used for the prediction of diabetes disease [12]. 91.7% of accuracy was accomplished by using this technique. In the detection of retinopathy diabetes, a deep neural network technique was embraced, through which 74.4% of accuracy was accomplished via CNN [13]. AI along with CNN classifier, a multiclass retinal diabetes detection was taken place through which 92% of accuracy was achieved.[14]. For the prediction of cardiovascular disease and diabetes, a data-driven approach was carried out along with machine learning correlated with a support vector machine, linear regression and random forest classifier. By using this technique,

95.7% of accuracy was accomplished [15]. An innovative methodology was carried out for the prediction of diabetes is based on smartphones [16]. For disease diagnoses, data of images was examined in this paper. For measuring the level of blood glucose in patients, an agent based on a microcontroller was used [17]. An integrated sensor therapy was used to observe glucose levels in patients with diabetes [18]. A smart application based on a self-recommendation system was introduced and trained for recording health data for example physical activities of patients and other essential specifications regarding diabetes [19]. For the prediction of diabetes type 2, a model of ensemble classification was carried out. 82.2% of accuracy was accomplished through this model [20]. The latest research has been conducted on the Detection of diabetes using machine learning algorithms, Ihnaini et al [21] achieves 99.6 accuracy after data fusion. Our experiment achieves 99% accuracy without data fusion.

**METHODOLOGY**

This work was carried out by searching relevant published papers available on IEEE, PubMed and Google scholar. These articles were retrieved by placing a search term in the relevant search engine “Diabetes Prediction through Data Mining”. This results in 11,800 published papers on Google scholar while 300 relevant results were shown on IEEE and only five relevant articles were found on Google Scholar.

Furthermore, we aimed to include the latest research. We excluded the repeated articles; 200 articles were considered. Further segregation was done in which articles with full-length availability and those that were in English were included. Only research papers and blog articles were included whereas data included in the book, letter to the editor, short communication and patent were excluded. Similarly, articles without any abstract were also not considered.

Table 2 given below shows the details of the process.

Table 2: Methodology

| Searched term "Diabetes Prediction Through DATA MINING" | Most relevant articles focusing on subject | Irrelevant, lack abstract, not full-length, not in English | Articles shown on IEEE |
|---|---|---|---|
| 11,800(Google Scholar) | 30 articles | 11,70 | 300 articles |

**Classification**

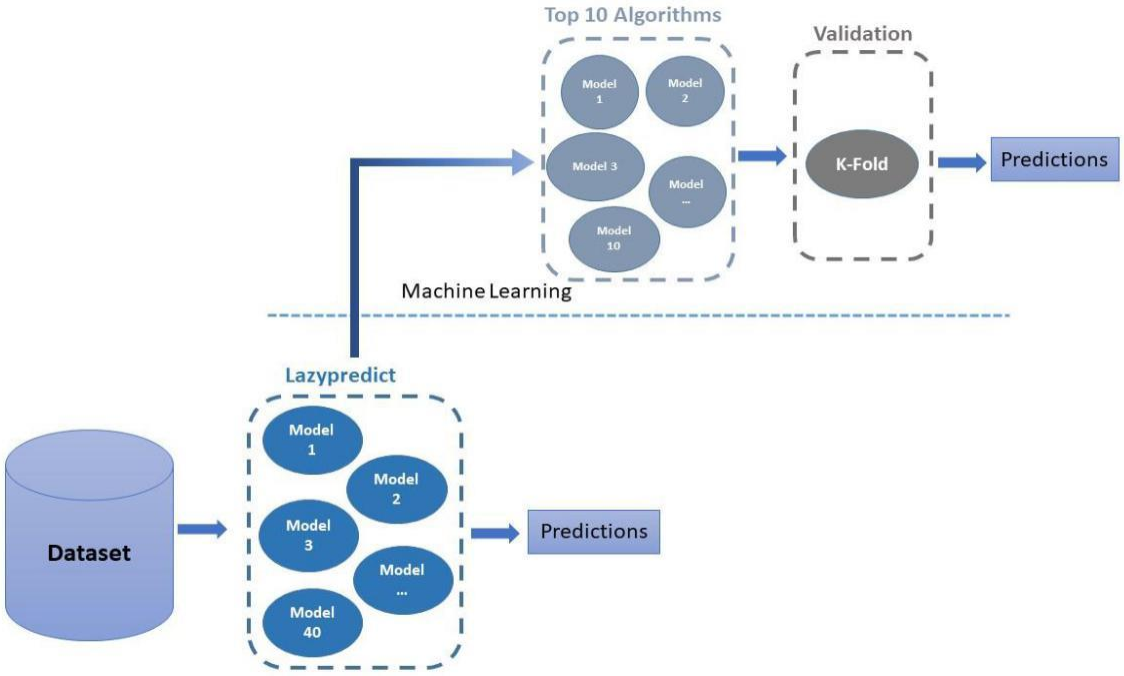
In this section, Lazypredict is used for classification. Lazypredict is an openly available python library developed by Shankar et al [22]. Lazypredict provides 40 machine learning algorithms for classification and regression. The data was divided into 2 different parts, training and testing using Sklearn. The Lazypredict classifier was trained on the training dataset and evaluated on the testing dataset.

Validation

K-Fold cross-validation is used to validate the machine learning models. The top 10 machine learning algorithms were selected and K-Fold cross-validation is performed on those classifiers. In K-Fold dataset is divided into k-subsets and each subset is used to evaluate the model. After that accuracies mean score is calculated. In this way, every data point is tested and used in training.

After the data collection machine learning algorithms were used for the experiment. The following diagram shows the architecture of the proposed methodology used in this research.

Figure 1 Architecture of proposed Methodology.




**DATASET DESCRIPTION**

By using electronic data, we have extracted beneficial information. Electronic medical data of patients is a challenging task. A dataset "PIMA INDIAN Diabetes Database" is collected from Indian based health organization and the other dataset on diabetes

is taken from the "Hospital Frankfurt Germany" electronically. These datasets were collected from Kaggle, dataset includes Frankfurt [7] dataset (D1) 2000 cases and PIMA [6] (D2) 768 cases with 9 attributes on basis of their results and values to target class which is diabetes. We utilized the same dataset because the data was accurate. All attributes are discussed and described in table 1 given below. Outcome is the target class, all the other attributes in the table help the algorithms to predict whether this person is diabetic or not. This is a two-class problem, in this class 0 is for No Diabetes and 1 is for Diabetes.

Table 1: Dataset Description

| Pregnancies | Glucose | BloodPress | SkinThickness | Insulin | BMI | Diabetes Pedigree Function | Age | Outcome |
|---|---|---|---|---|---|---|---|---|
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

The above table describes the type of data used in this research, Pregnancies here mean no. of times a female got conceived whereas skin thickness here means millimeters of skin which have to be pierced to give insulin. However pedigree here means people living in specific geographical areas like that Caucasians, Africans and others. Body mass index is calculated by dividing the mass (in KGs) by height (in meter squares). The normal range is between 18.5 to 24.5. Where glucose and blood pressure normal range varies from age to age.

**RESULTS**

The dataset is split into 80% training and 20% for testing ratios. 10 Algorithms are applied to the dataset and their accuracy is measured. After that K-Fold Algorithm is applied to the dataset. In K-Fold Algorithm, the dataset is divided into 10 partitions to train and test the model to measure the cross-validation score, after that means, and the standard deviation is calculated to measure the actual accuracy of the models. Results are in Table 3.

Table 3: Results

| Classifier | Accuracy for Data | | K-Fold N_Splits=10 For 2000 | | K-Fold N_Splits=10 For 768 | | Time Taken (Seconds) |
|---|---|---|---|---|---|---|---|
| | 2000 Tuples (D1) | 768 Tuples (D2) | Mean | Standard Deviation | Mean | Standard Deviation | |
| LGBM | 99 % | 69 % | 97.25 % | 1.62 % | 75.41 % | 5.12 % | 0.12 |
| XGB | 99 % | 69 % | 96.94 % | 1.73 % | 75.58 % | 6.45 % | 0.22 |
| Random Forest | 99 % | 98 % | 97.38 % | 1.50 % | 97.31 % | 1.27 % | 0.29 |
| Extra Trees | 99 % | 75 % | 97.06 % | 1.64 % | 75.09 % | 6.90 % | 0.57 |
| Decision Tree | 98 % | 75 % | 96.75 % | 1.74 % | 70.53 % | 6.30 % | 0.07 |
| Label Propagation | 98 % | 65 % | 96.13 % | 1.48 % | 66.28 % | 5.93 % | 0.24 |
| Label Spreading | 98 % | 65 % | 96.13 % | 1.48 % | 66.28 % | 5.93 % | 0.37 |
| Extra Tree | 97 % | 67 % | 96.50 % | 1.80 % | 66.77 % | 5.91 % | 0.03 |
| Bagging | 97 % | 72 % | 96.06 % | 1.61 % | 75.73 % | 5.45 % | 0.15 |
| KNeighbors | 79 % | 66 % | 78.94 % | 4.09 % | 70.67 % | 8.62 % | 0.07 |

After training and testing all the models, LGBM (Light Gradient Boosting Machine) achieves the highest accuracy of 99% with a minimum time of 0.12 seconds D1. LGBM handles large size of data, and it takes lower memory to run. KNeighbors takes less time than LGBM but accuracy is lower than LGBM. The models were tested on two different datasets, D2 but its accuracy is lower, but when the data size is increased its accuracy improved. 10 Fold cross-validation testing is applied to the ML algorithm to equally test all the portions of the dataset. Because Machine learning algorithms required a large amount of data to train the model for all the possible scenarios. If the dataset is small, it results in poor approximation. And if there is a multiclass problem then results can give slight variation.

**COMPARATIVE ANALYSIS**

Ihnaini et al [21] haven't used multiple python algorithms to solve the problem. They purposed their deep ensemble learning model with an accuracy of 72.73 % for 768 tuples and 91.00 % for 2000 tuples before data fusion. After data fusion of D1 and D2, their model achieves an accuracy of 99.6 %. It failed to compete with LGBM, XGB, Random Forest and ExtraTrees algorithms because their accuracy before data fusion is 99 %. Comparison results are in Table 4. First, their scope was limited, our research was conducted with 40 machine learning algorithms. 10 Algorithms were selected based on their accuracy and the time they have taken. After that, we performed K-Fold cross-validation testing to get better and more accurate results. By using K-Fold cross-validation testing the dataset is divided into 10 different parts and with all these parts training and testing have been done. On the other hand, python libraries can solve this problem without data fusion.

Table 4: Ihnaini et al [21]

| Classifier | Accuracy (%) for 768 Tuples | Accuracy (%) for 2000 |
|---|---|---|
| Logistics Regression | 74.68 | 77.75 |
| Naïve Bayes | 72.08 | 76.50 |
| Random Forest | 74.68 | 81.25 |
| K-nearest Neighbor | 73.38 | 77.75 |
| Decision Tree | 74.03 | 83.75 |
| Support Vector Machine | 74.68 | 84.00 |
| Purposed Model | 72.73 | 91.00 |

**CODE & DATASET**

We have used 2 datasets collected from 2 different databases. The first 768 tuples were collected from PIMA Indian diabetes database [6] and the second 2000 tuples were collected from the Frankfurt Germany hospital database [7]. After that we applied 10 Different machine learning algorithms to check which algorithm provided the highest accuracy with the minimum time including K-Fold Cross validation testing, Code is available at Github [8].

**Conclusion & Future Research Direction**

We have developed an intelligent health care system to detect diabetes using machine learning algorithms. 40 machine learning algorithms were used along with K-Fold cross-validation testing for each algorithm, top 10 algorithms are selected based on their accuracy and the time they took. LGBM classifier has achieved the maximum accuracy of 99% with a minimum time of 12 Milliseconds compared with the other classification algorithms, after conducting K-

Fold Cross validation testing it achieves the accuracy of 97.25% mean and 1.62% standard deviation. The following algorithms can be used in hospitals and laboratories, where if a patient came the system can tell whether this person is diabetic or not, or else the system can recommend performing a diabetes test. The code used in this research is made public, anyone can download that code and use this as per their requirement. We have made the code available on Github for future research direction. You just have to import the dataset, chose your target class and can run the code simply on GoogleColab and Anaconda JupyterLab with K-Fold Cross Validation testing for all the top 10 Algorithms. The code can also be used for Protein, DNA and RNA sequences. Zhen et al

[23] created a live server for protein and peptides sequences feature extraction. First, we have to extract the features from the sequence and then those extracted features can be passed to those classifiers to make predictions.

**Authors Contribution**

**Saman Liaqat:** Conceptualization, Methodology, Software. **Allah Dita, Aneela Mehmood, Malik Tahir Hassan**: Data curation, Writing- Original draft preparation. **Hassan Kaleem**: Visualization, Investigation, Validation, Writing-Reviewing and Editing

**Compliance with Ethical Standards:**

It is declared that all authors don't have any conflict of interest. It is also declared that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

**REFERENCES**


[1] "Global report on diabetes." https://www.who.int/publications/i/item/9789241565257 (accessed Feb. 09, 2022).

[2] Home *et al.*, "Resources | IDF Diabetes Atlas." https://diabetesatlas.org/resources/ (accessed Feb. 09, 2022).

[3] "Prof. Min Chen's Home Page." https://people.ece.ubc.ca/minchen/ (accessed Feb. 09, 2022).

[4] S. Afzali and O. Yildiz, "An Effective Sample Preparation Method for Diabetes Prediction," *Int. Arab J. Inf. Technol.*, vol. 15, pp. 968–973, Nov. 2018.

[5] N. S. Artzi *et al.*, "Prediction of gestational diabetes based on nationwide electronic health records," *Nat. Med.*, vol. 26, no. 1, pp. 71–76, Jan. 2020, doi: 10.1038/s41591-019-0724-8.

[6] "Pima Indians Diabetes Database." https://kaggle.com/uciml/pima-indians-diabetes-database (accessed Feb. 12, 2022).

[7] "diabetes." https://kaggle.com/johndasilva/diabetes (accessed Feb. 12, 2022).

[8] RaoHassanKaleem, *RaoHassanKaleem/Diebetes-Detection-using-Machine-Learning-Algorithms*. 2022. Accessed: Feb. 14, 2022. [Online]. Available: https://github.com/RaoHassanKaleem/Diebetes-Detection-using-Machine-Learning-Algorithms

[9] R. Aminah and A. H. Saputro, "Diabetes prediction system based on iridology using machine learning:6thInternational ConferenceonInformation Technology, Computer and Electrical Engineering, ICITACEE 2019," *2019 6th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2019*, Sep. 2019, doi: 10.1109/ICITACEE.2019.8904125.

[10] A. Mohebbi, T. B. Aradottir, A. R. Johansen, H. Bengtsson, M. Fraccaro, and M. Morup, "A deep learning approach to adherence detection for type 2 diabetics," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2017, pp. 2896–2899, Jul. 2017, doi: 10.1109/EMBC.2017.8037462.

[11] "Analysis of diabetes mellitus for early prediction using optimal features selection | Journal of Big Data | Full Text." https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6 (accessed Jan. 20, 2022).

[12] B. Ihnaini *et al.*, "A Smart Healthcare Recommendation System for Multidisciplinary Diabetes Patients with Data Fusion Based on Deep Ensemble Learning," *Comput. Intell. Neurosci.*, vol. 2021, p. e4243700, Sep. 2021, doi: 10.1155/2021/4243700.

[13] "Detection of Multi-Class Retinal Diseases Using Artificial Intelligence: An Expeditious Learning Using Deep CNN with Minimal Data | Biomedical and Pharmacology Journal." https://biomedpharmajournal.org/vol12no3/detection-of-multi-class-retinal-diseases-using-artificial-intelligence-an-expeditious-learning-using-deep-cnn-with-minimal-data/ (accessed Jan. 21, 2022).

[14] "Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy | Eye." https://www.nature.com/articles/s41433-018-0269-y (accessed Jan. 21, 2022).

[15] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 211, Nov. 2019, doi: 10.1186/s12911-019-0918-5.

[16] R. Rajalakshmi, R. Subashini, R. M. Anjana, and V. Mohan, "Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence," *Eye*, vol. 32, no. 6, pp. 1138–1144, Jun. 2018, doi: 10.1038/s41433-018-0064-9.

[17] M. Chen, J. Yang, J. Zhou, Y. Hao, J. Zhang, and C.-H. Youn, "5G-SmartDiabetes:Toward

Personalized Diabetes Diagnosis with Healthcare Big Data Clouds," *IEEE Commun. Mag.*, vol. 56, pp. 16–23, Apr. 2018, doi: 10.1109/MCOM.2018.1700788.

[18] P. Choudhary, S. de Portu, A. Arrieta, J. Castañeda, and F. M. Campbell, "Use of sensor-integrated pump therapy to reduce hypoglycaemia in people with Type 1 diabetes: a real-world study in the UK," *Diabet. Med. J. Br. Diabet. Assoc.*, vol. 36, no. 9, pp. 1100–1108, Sep. 2019, doi: 10.1111/dme.14043.

[19] A. Steinert, M. Haesner, and E. Steinhagen-Thiessen, "App-basiertes Selbstmonitoring bei Typ-2-Diabetes," *Z. Für Gerontol. Geriatr.*, vol. 50, no. 6, pp. 516–523, Aug. 2017, doi: 10.1007/s00391-016-1082-5.

[20] B. P. Nguyen *et al.*, "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records," *Comput. Methods Programs Biomed.*, vol. 182, p. 105055, Dec. 2019, doi: 10.1016/j.cmpb.2019.105055.

[21] B. Ihnaini *et al.*, "A Smart Healthcare Recommendation System for Multidisciplinary Diabetes Patients with Data Fusion Based on Deep Ensemble Learning," *Comput. Intell. Neurosci.*, vol. 2021, p. e4243700, Sep. 2021, doi: 10.1155/2021/4243700.

[22] "Welcome to Lazy Predict's documentation! — Lazy Predict 0.2.9 documentation." https://lazypredict.readthedocs.io/en/latest/ (accessed Jun. 15, 2022).

[23] Z. Chen *et al.*, "iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018, doi: 10.1093/bioinformatics/bty140.