IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306183|

Enhancing Weather Forecast Accuracy through the Implementation of Mode based Data Mining Technique

Manpreet Kaur, Mohit Trehan

Department of Computer Science, GCET, Gurdaspur, India

ABSTRACT: Weather prediction is a multifaceted application requiring several phases of processing. The proposed mechanism employs mode-based pre-processing combined with a clustering approach to enhance forecast accuracy. Initially, the dataset is loaded and pre-processed to handle outliers and missing values using a mode-based strategy. Correlation and clustering techniques are then applied for labeling information. Clustering allows for the prediction of multiple classes, achieving a classification accuracy of approximately 80%, which is a 4% improvement over existing methods. This mechanism effectively manages noise, enhancing classification accuracy. The dataset for this study is sourced from Kaggle. To address noise, a mean-based pre-processing mechanism is used, replacing noisy data with the column mean. After pre-processing, segmentation retains critical dataset parts while removing non-critical sections. Classification is then performed, further improving accuracy through the combined mode-based approach and pre-processing. Modern transportation systems face challenges due to increasingly complex road weather conditions. Intelligent Weather Systems have been developed to tackle these issues, providing accurate and hassle-free forecasts crucial for safe and smooth driving. These systems have been a focus of continuous research and development, with significant advancements in weather prediction using IoT-generated datasets. Techniques like weather flow analysis have been employed to address forecasting challenges. This paper provides a comprehensive analysis of previous research in weather prediction, highlights significant advancements, and proposes future research directions. Overall, the proposed mechanism effectively improves weather prediction accuracy and contributes to the ongoing progress in this field.

I. INTRODUCTION

1.1 Data Mining

Mining indicates the process of filtering the data. Volume of data risen substantially over the period. Extracting meaningful information from such vast amount of information could be tedious task. To accomplish the filtering role of data mining is crucial. This work explores the features of data mining to predict weather. The steps associated with mining are elaborated in the 1.2. The flow of data mining process is given within the figure 1

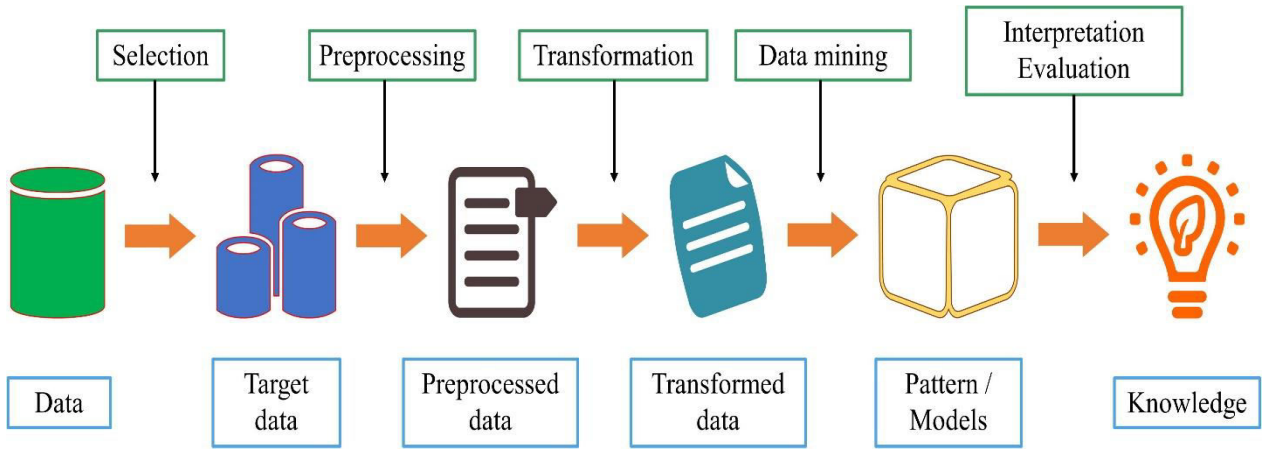

Figure 1: Mining process

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306183|

1.2 Phases associated with Mining

The entire process of data mining is separated into phases. Phase wise methodology diminishes the complexity of operation. Each phase contributes significantly towards achieving best possible classification accuracy associated with assigned task. These phases are elaborated as under

1.2.1 Data Acquisition or Selection

This is one of the most crucial phases associated within mining. In this phase relevant data is collected either through benchmark websites or with real time mechanism. The formed dataset is generally in the form of CSV. Sometimes dataset is not in proper format and must be brought into relevant form using data analysis process available in Microsoft suite. After eliminating unnecessary fields, target dataset is formed.

1.2.2 Pre-processing

The target dataset can include noise. Outliers or missing values are two examples of this noise. Missing values indicates columns that have missing cells. These missing cells impact final classification. To eliminate missing values, mean, median or mode-based approach can be followed. Outlier indicates the abnormal values from the cells of the dataset. Outliers can be eliminated using box layout-based mechanism. After handling noise from the dataset, classification becomes simplified.

1.2.3 Transformation

The dataset contains values that could be large and hence calculations become complex. This overall detection rate thus becomes poor. To overcome the issue, normalization is performed. The normalization indicates the division of values of each cell with the maximum value of each column. This will reduce the magnitude of values within the dataset from '0' to '1'.

1.2.4 Data Mining

Mining procedure will filter the data towards the desired class. The class will be the prediction result by analyzing the metrics present in the dataset. Each attribute within the dataset is termed as feature. To perform classification, each feature within the dataset is analyzed and final classification is made.

1.2.5Interpretation

Final interpretation is the classification of test data within particularly class. The degree of misleading or misclassified values decreases the classification accuracy.

The overall process of mining is implemented within the proposed work. The proposed work consists of using ensemble-based approach for prediction of weather.

II. LITERATURE SURVEY

This section provides in depth analysis of techniques used for identifying and predicting weather conditions.

Fairoz Q. Kareem1(2023)[37] stated that with the recent development in technology, especially in the DM and machine learning techniques, many researchers proposed weather forecasting prediction systems based on data mining classification techniques. In this paper, we utilized neural networks, Naïve Bayes, random forest, and K-nearest neighbor algorithms to build weather forecasting prediction models. These models classify the unseen data instances to multiple class rain, fog, partly-cloudy day, clear-day and cloudy. These model performance for each algorithm has been trained and tested using synoptic data from the Kaggle website. This dataset contains (1796) instances and (8) attributes in our possession. Comparing with other algorithms, the Random forest algorithm achieved the best performance accuracy of 89%. These results indicate the ability of data mining classification algorithms to present optimal tools to predict weather forecasting.

Krishnaveni (2022)[19] discussed a weather prediction using sprint algorithm. This approach is used to analyze the different patterns that originate from the dataset. Train and test split method is used for the prediction of weather. The result of prediction is presented through classification accuracy. The classification accuracy achieved through the approach is less then 90%.

J., (2020)[16] proposed a numerical based mechanism to predict the weather. The prediction is based on training using supervised learning mechanism. The size of the dataset is limited in nature. The rate of detection on this dataset is fast.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306183|

However large dataset is not tested through this approach. The parameter used for evaluation includes classification accuracy which is less then 80%.

N. et al., (2020)[27] suggested a wind power prediction using gaussian processes. The numerical method is used for the prediction of the weather. The predicted weather can be used by the clients for making decision regarding going out. The medical emergencies can also be affected through this prediction. Result of the prediction is generally expressed through classification accuracy.

R. et al., (2022)[30]proposed a wind speed forecasting model using neural network-based approach. This approach is a layered based approach. In the first layer, data accumulation is established. Second layer is processing layer that is used to extract the decision using training dataset. Output layer is used for the final prediction regarding weather. The classification prediction within the range of 90% is achieved.

Sheikh et al., (2020)[33] stated various analysis regarding data mining strategies for weather forecasting. Different classification approaches including KNN, SVM , Random Forest and naïve Bayes algorithms are applied on the dataset to extract the result of prediction. Result is expressed through classification accuracy.

Kareem (2022)[18]discussed the weather forecasting through the data mining approach. The data mining approaches including KNN, SVM, Naïve Bayes and decision tree are used for classification. The classification accuracy of support vector machine appears to be within 80 to 85%.

Dhamodaran (2023)[9]proposed weather prediction using random forest-based approach. The GIS dataset that is large can be easily categorized using this approach. The mechanism provided better classification accuracy. The classification accuracy within range range of 80% is achieved through this approach.

Wu Zhu (2019)[36]proposed a naïve bayes based approach for the detection and prediction of weather. The bayes theorem is followed during prediction. This mechanism does not use any optimization. The classification accuracy is thus poor in this case. The mechanism ensures that different classes of weather can be predicted using this mechanism.

Singh Chaturvedi (2019)[34] implemented a machine learning based mechanism for predicting weather. The weather forecasting is divided into 4 different classes. The classification accuracy using this approach is high but large dataset may not be tackled using this approach. The classification accuracy could be further improved by applying median or mode-based approach.

Madan et al., (2018) [22]discussed the weather forecasting on big data. Size of data keeps on growing and extracting meaningful information from such volume of data could be an issue. This issue is tackled within this work. Hadoop is used for the processing of big data. Prediction rate is however minimal but provide good classification accuracy through this approach.

III. PROBLEM DEFINITION

The ongoing model utilizes brain organizations, Nave Bayes, arbitrary backwoods, and K-closest neighbor to make expectations. Because no connection between data is assessed, all data, with or without a substantial effect on prediction, must be included in the analysis. Weather forecast performance can be improved by incorporating a correlation-based mechanism into the prediction model. In the existing system, there is no outlier detection and a weak pre-processing method for missing values handling. The optimization techniques like NN and SVM are not used in the clustering and deep learning mechanisms used in numerous literatures. Only one illness level is detected as a result. The created feature vector will be used to forecast disease when compared to the test dataset features. If the hyper plane is not pierced, the detection, while potentially accurate, may also present a problem. Another difficulty is the amount of the dataset. For deep learning to extract enough features, a rather big dataset is required. If the dataset is too small, classification accuracy will suffer. NN must be a part of the classification layer of deep learning for the early disease detection and processing. In the event when two hyper planes were used, the classification accuracy was also compromised.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306183|

The existing method achieves a classification accuracy of 89 percent. Using a correlation method, this can be improved even more. The issue with the existing approach is given as under

- Classification accuracy with the existing approach can be improved further
- Pre-processing mechanism can be improved using regression-based modeling.
- Analyzing data without correlation lead to poor detection rates.

3.1 Objectives of Study

The objectives associated with the detection weather are listed in this section. The objective of study indicates what is to be achieved through the proposed model.To improve the pre-processing phase by incorporating an elbow mechanism to detect outliers.

- To apply the train test split mechanism for achieving training and testing dataset
- To increase prediction performance using correlation-based analysis
- To compare the accuracy of existing and prospective work in terms of classification accuracy, F-Score, Sensitivity and specificity

IV. PERFORMANCE ANALYSIS AND RESULTS

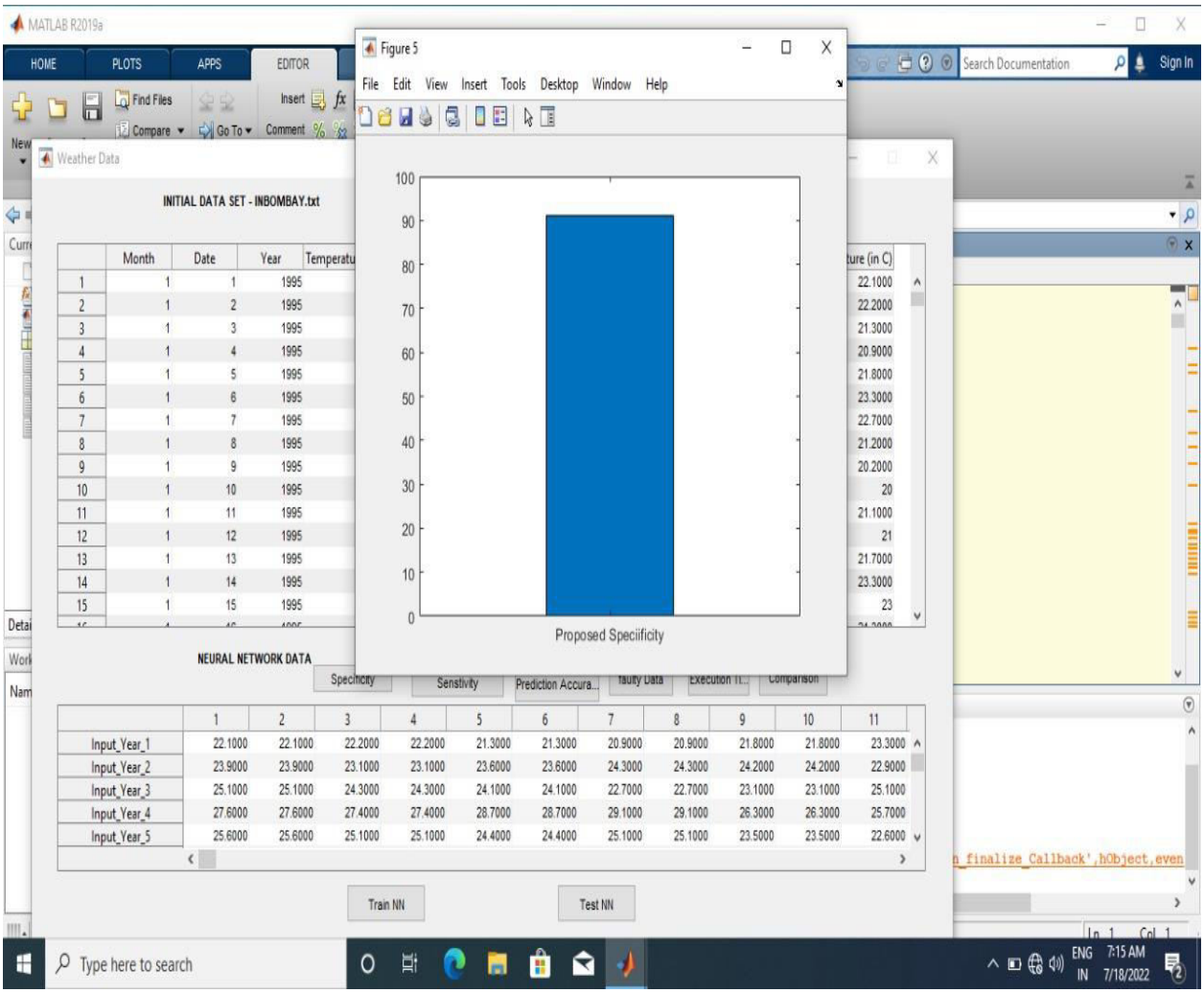

Figure2: Proposed Specificity

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306183|

The next result is in terms of classification accuracy. This means that both true positive and true negative values predicted with the proposed work is high. This means that classification accuracy proves the worth of study.

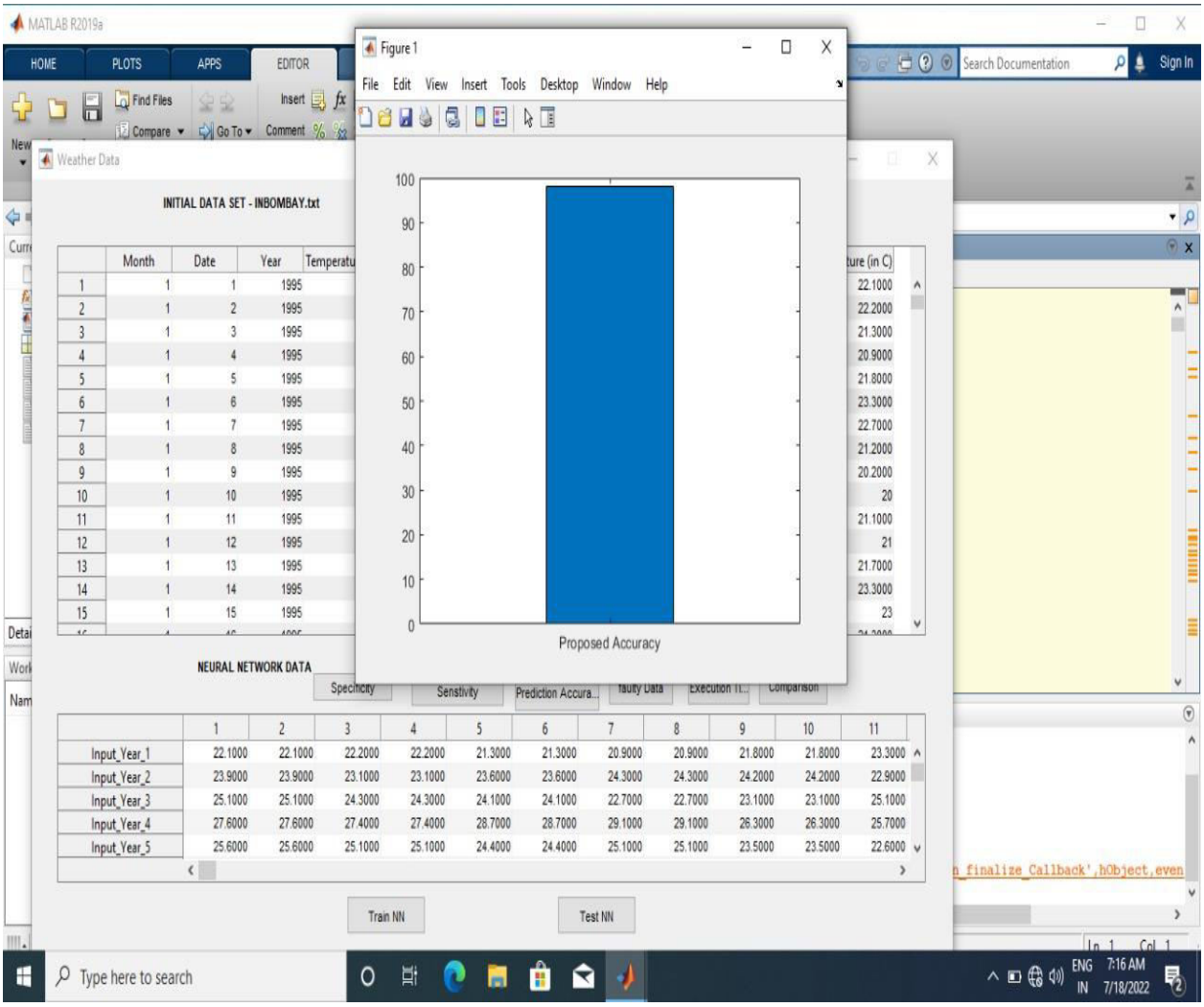

Figure 3: Classification accuracy with the proposed approach

Next comparison is in the form of execution time. Execution time with the proposed approach is reduced considerably. This is primarily due to reduce size of the dataset. Correlation based approach will be used to obtain the dataset with relevant fields and using meaningful information only with the proposed approach

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306183|

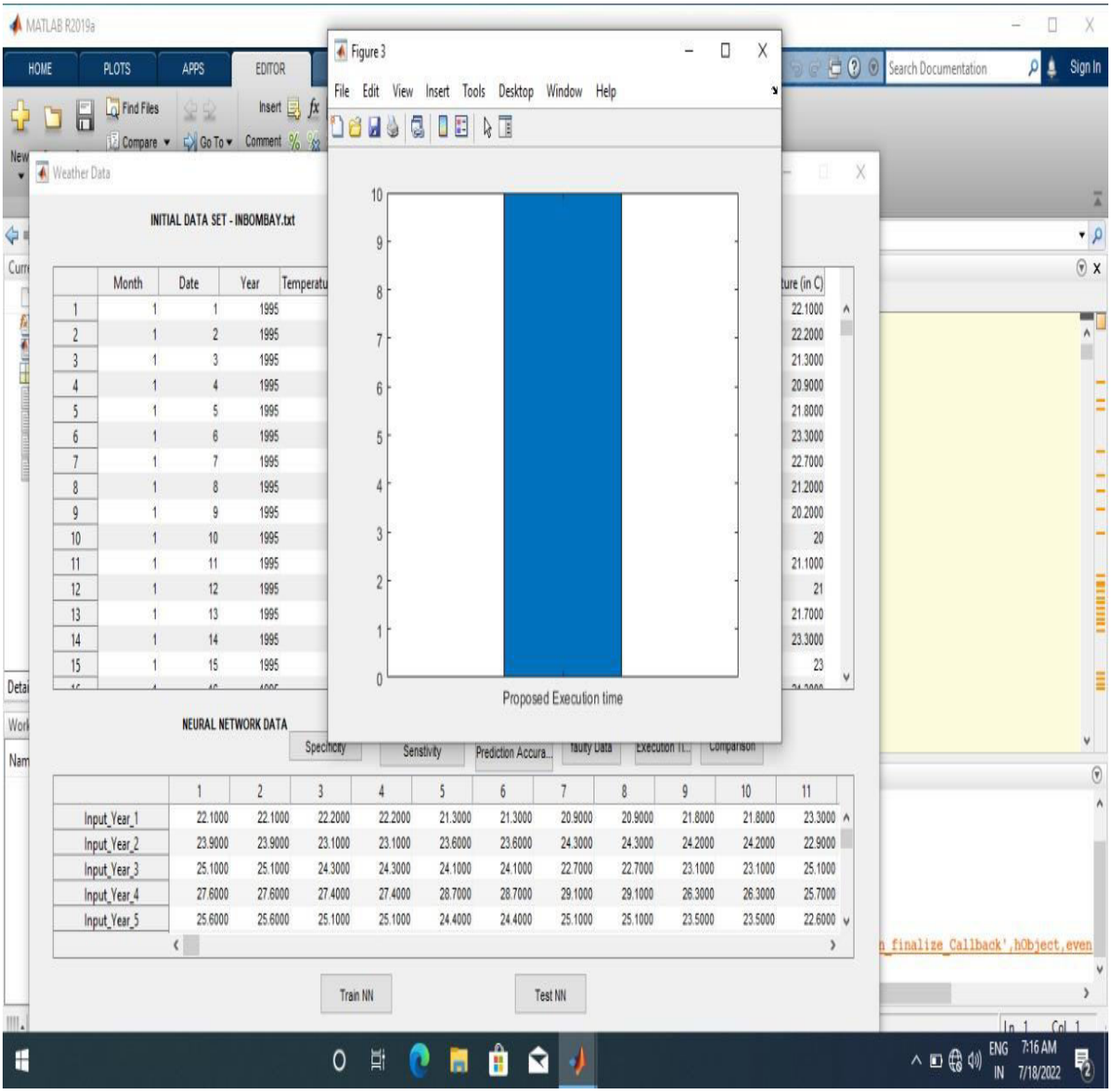

Figure4: Execution time with the proposed approach

Next comparisons indicate the difference of existing and proposed approach with different metrics including classification accuracy, specificity, sensitivity and execution time. Figure 14 indicates the comparison of first metric.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306183|

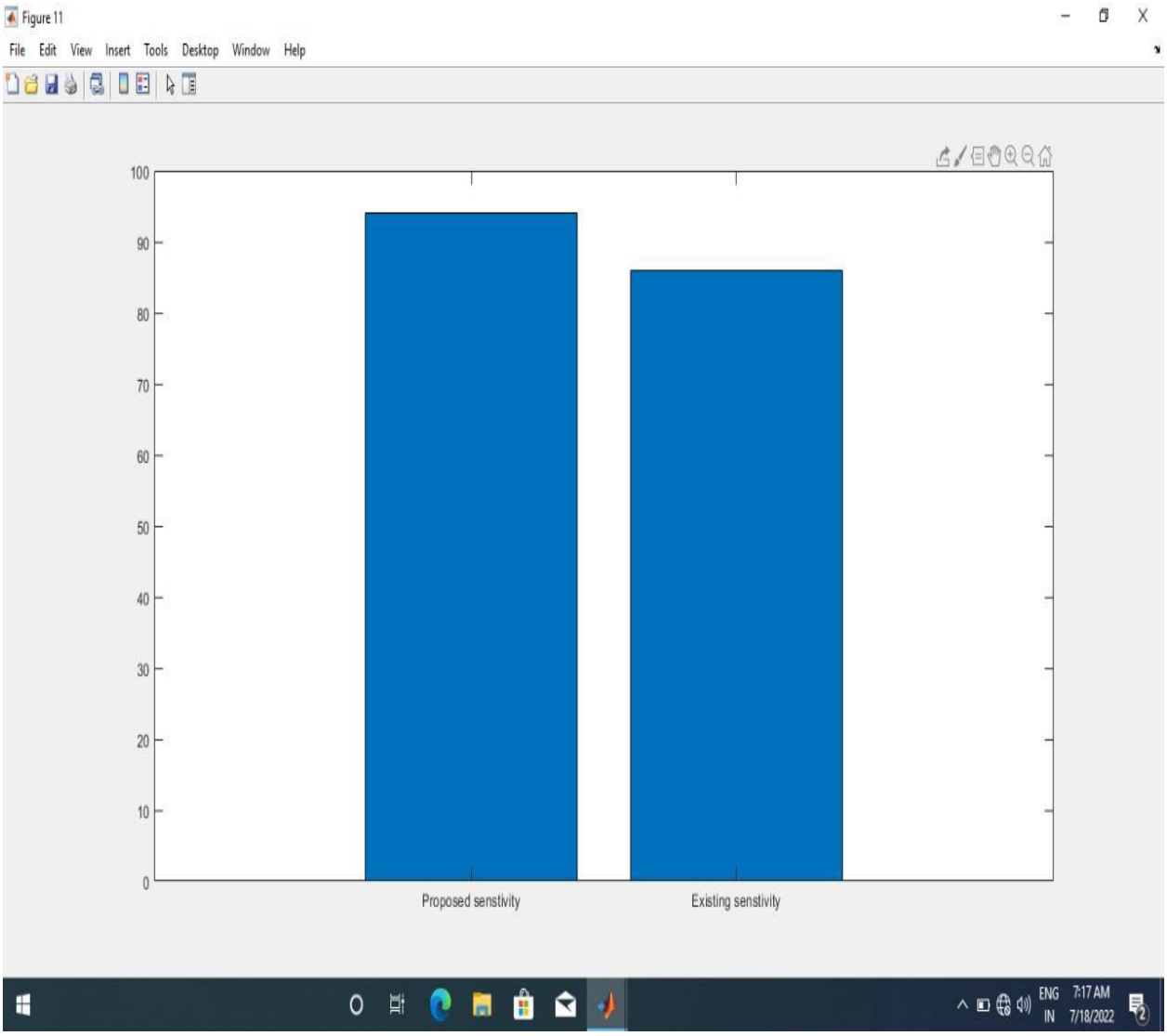

Figure5: Sensitivity comparison

The sensitivity indicates the degree of true negative values predicted with the proposed and existing approach. With proposed approach sensitivity is reduced and hence proves the worth of study.

V. CONCLUSION AND FUTURE SCOPE

Conclusion

The proposed work uses the correlation based neural network-based approach for gathering the meaningful information from the dataset. The pre-processing-based mechanism first of all ensures that noise from the dataset is eliminated. The noise from the dataset is also termed as missing values and extreme values from the dataset is termed as outliers. The elimination of both from the dataset ensures that better classification accuracy is generated at the end. The entire implementation is carried out within Matlab and graphical user interface is used. The graphical user interface ensures that better classification accuracy is achieved. After pre-processing, result will be displayed within the first section of the screen. The mechanism also improves the classification accuracy. After this size of the data will be reduced. The primary reason for the same is improving execution time. Higher correlation values will be retained and lower correlation values will be rejected. Once the dataset is initialized, it will be loaded within the GUI. The dataset will be loaded with the attributes represented distinctly. The mechanism also ensures that layers of neural network can be used

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306183|

easily. Furthermore, neural network layer is used to process the data from the dataset. The optimization-based approach is applied to achieve the better output. The overall process is repeated until the desired result is obtained.

Future Scope

The entire process of weather prediction was applied over the text dataset that is of smaller size. The smaller size dataset produced optimal results. However in future same mechanism can be tested over the large dataset. The large dataset may show deviation within the results. The pre-processing mechanism can be improved further using the optimized classifiers with spider monkey approach. The execution time however can be increased but better classification accuracy can be achieved with the spider monkey approach.

REFERENCES

1. Adarsh (2013) "Multiclass svm-based automated diagnosis of diabetic retinopathy," International Conference on Communications and Signal Processing, pp. 206–210. Available at: https://doi.org/10.1109/iccsp.2013.6577044.
2. Ahuja, K. (no date) "Data Elimination Based Technique for Mining Frequent Closed Item Set," pp. 1–4.
3. Al-ayyoub, M. et al. (2016) "A GPU-Based Breast Cancer Detection System Using Single Pass Fuzzy C-Means Clustering Algorithm," IEEE [Preprint].
4. Algwaiz (2016) "Data Mining Using Probabilistic Grammars," IEEE Access, pp. 1–4.
5. Author, Z.B. (2016) "Data-mining behavioral data from the web," pp. 122–127. Available at: https://doi.org/10.1109/SKIMA.2016.7916208.
6. Bakon, M. et al. (2017) "A Data Mining Approach for Multivariate Outlier Detection in Postprocessing of Multitemporal In SAR Results," IEEE Access, 1, pp. 1–8.
7. Brown (2017) "Improving Privacy Preserving Methods to Enhance Data Mining for Correlation Research," pp. 3–6.
8. Bulsara, V. et al. (2021) "Low Cost Medical Image Processing System for Rural / Semi Urban Healthcare," IEEE Access, pp. 724–728.
9. Dhamodaran,(2023) "Weather prediction model using random forest algorithm and GIS data model," Lecture Notes on Data Engineering and Communications Technologies, 46, pp. 306–311. Available at: https://doi.org/10.1007/978-3-030-38040-3_35.
10. Doshi, D. et al. (2016) "Diabetic Retinopathy Detection using Deep Convolutional Neural Networks," 2016 International Conference on Computing, Analytics and Security Trends [Preprint].
11. Dutta, S. et al. (2018) "Classification of diabetic retinopathy images by using deep learning models," International Journal of Grid and Distributed Computing, 11(1), pp. 89–106. Available at: https://doi.org/10.14257/ijgdc.2018.11.1.09.
12. Efremov, V.P. et al. (2017) "Deep Learning for Health Informatics," Journal of Experimental and Theoretical Physics Letters, 81(8), pp. 378–382. Available at: https://doi.org/10.1109/JBHI.2016.2636665.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details