BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Exploring the association between stroke and weather conditions using Artificial Intelligence

- Research Project -

Ploscar Andreea Alina

2022-2023

Contents

1 Abstract

2 Introduction
2.1 Aim
2.2 Related Work
2.3 Original Contribution

3 Datasets
3.1 Meteorological Data
3.2 Medical Data

4 Methods
4.1 Meteorological Data Visualization
4.2 Medical data Clustering
4.2.1 K-Means Clustering
4.2.2 Mean-Shift Clustering
4.3 Identifying Variations
4.3.1 Feature Importance
4.3.2 Outliers Identification
4.4 Neural Network

5 Experimental Modelling
5.1 Data
5.2 Metrics
5.3 Formal Model

6 Case study
6.1 Data processing
6.2 Data visualization
6.2.1 Meteorological Data
6.3 Data labeling
6.4 Experiment and Results

7 Related Work - Comparison
7.1 Real Dataset
7.2 Approaches and Metrics

8 Ethics

9 Discussion

10 Conclusion

References

1 Abstract

The study presents a new approach for assessing the impact of weather conditions on stroke incidence, making use of artificial intelligence to find this correlation and predict a sudden increase in the incidence of strokes based on meteorological conditions such as differences in atmospheric pressure, temperature, atmospheric fronts.

2 Introduction

Stroke was the second cause of death in Romania in 2016 [Don18] and is a major cause of mortality worldwide. Patients hospitalized with stroke can be treated, but due the sudden appearance of stroke symptoms, these patients require immediate attention from healthcare professionals, as well as available operating rooms. This process would be more efficient if healthcare workers could be aware that the incidence of strokes in the area is expected to increase, thus they would organize the non-emergency cases and the operating rooms using this information.

2.1 Aim

The aim of the study is to assess the correlation between weather conditions and stroke incidence in Transylvania, Romania. As weather conditions temperature, air pressure, atmospheric fronts, precipitations and differences in these values are taken into consideration when analysing the correlation. For a better understanding of the relationship, underlying conditions of patients are taken into consideration. As a limitation, the study does not assess the impact of personal stress factors on patients hospitalized with stroke.

2.2 Related Work

- "A study of weekly and seasonal variation of stroke onset"

  The paper [WSK02] is relevant to this study as it presents the relationship between seasons, week days, age and stroke incidents. The results show a significant weekly and seasonal variation in the occurrence of stroke and a negative dose response relationship between seasonal variations in occurrence and age. This may be caused by the significant impact that lifestyle has on the probability of a stroke, topic that will be further discussed in this study.

  Structure:

  - Abstract - short overview of the article, containing the problem description, the time interval, the geographical area where the study was conducted, the methods and the results
  - Introduction - presents results and limitations of existing studies and the aim of the study
  - Subjects and methods - geographical and meteorological details of the targeted area, data collection and statistical methods used
  - Results
    - population characteristics
    - weekly variation of stroke occurrence and the effect of age
    - seasonal variation of stroke occurrence and the effect of age
  - Discussion - compares the obtained results to other studies, presents conflicting results and gives possible explanations for them, conclusion
  - References - list of cited and related work

  In this paper, references appear as a list in the last section, in the following format: authors (year) title journal volume:pages. When cited, the format is: (authors year).

  Citations: 105
  References: 73

- "Weather as a Trigger of Stroke"

  The paper [Jim+08] analyses the relationship between daily meteorological conditions and daily as well as seasonal stroke incidence. The approach is closed to our study, as it takes into consideration daily meteorological conditions, not only seasonal ones. The results show little association between atmospheric pressure(AP) and stroke, but higher association between stroke and variations in atmospheric pressure. The variation was computed as the value of AP compared to the previous day. Our study will also take into consideration 3, 6 and 12 hours variations.

  Structure:

  - Abstract - background, methods, results, conclusions
  - Introduction - presents results and limitations of existing studies and the aim of the study
  - Methods
    - Subjects - geographical, temporal and quantitative details of analysed data
    - Classification and Variables - classification of patients and types of strokes, recorded medical and meteorological data
    - Statistical Analysis - statistical methods used to analyse the data
    - Ethics - ethical guidelines
  - Results
    - Descriptive Data
    - Daily Incidence Analysis
    - Seasonal Analysis
  - Discussion - compares the obtained results to other studies, presents strengths and limitations of the study, conclusion
  - Acknowledgements - collaborations
  - References - numbered list of cited and related work

  In this paper, references appear as a numbered list in the last section, in the following format: authors title journal year;volume:pages. When cited, the format is: [number] representing index in references list.

  Citations: 73
  References: 34

- "The association between weather conditions and stroke admissions in Turkey"

  [Cev+15] is focused on data from turkey, area that is closer to the one covered by the present study, i.e. Romania, so the meteorological conditions will be more similar to the ones in our study. The paper takes into consideration ischemic stroke (IS), hemorrhagic stroke (HS) and subarachnoidal hemorrhage (SAH) and The results from this paper present no association between incidence of overall admissions due to strokes and meteorological parameters, but they do demonstrate an association between admissions due to SAH ans HS and weather conditions, especially temperature.

  Structure:

  - Abstract - aim of the study, details about used data, results
  - Introduction - presents results and limitations of existing studies and the aim of the study
  - Materials and Methods
    - Study Design - geographical, temporal and quantitative details of analysed data, diagnosis details
    - Meteorological data - geographical and meteorological details about data
    - Statistical Analysis - statistical methods used to analyse the data
  - Results - explained results

- Discussion - compares the obtained results to other studies, presents strengths and limitations of the study, conclusion
- References - list of cited and related work

In this paper, references appear as a list in the last section, in the following format: authors (year) title journal volume:pages. When cited, the format is: (author year)

Citations: 24
References: 22

- "An Improved Back Propagation Neural Network Model and Its Application"

The article [Li+14] approaches the problem of finding a relationship between stroke incidence and weather conditions using a back propagation neural network, method that is closer to the one further presented in the present study. The results are in line with the ones presented above, presenting a stronger relationship between stroke incidence and atmospheric pressure, and a weaker, negative relationship between stroke incidence and temperature.

Structure:

- Abstract - overview, aim of the study, details about used data, results
- Introduction - background, short description about methods
- Model imporvement
  - Notations - legend of notations and their meanings
  - The BPNN Flow - figure of the BPNN presenting the layers, input, output
  - Forward Propagation Process of Signal - mathematical explanations for hidden and output layers
  - Back Propagation Process of Error - mathematical explanation of used formulas and functions
  - The Improved BPNN Algorithm - description of improvements
  - Other Network Parameters - list of parameters for BPNN and environment
  - Model Solution - details about training samples input, samples training, samples prediction and effect, results
  - Model evaluation and promotion - describing the model as a prediction method
- Acknowledgments
- References - numbered list of cited and related work

In this paper, references appear as a list in the last section, in the following format: authors, title, journal, volume, no, pages, year. When cited, the format is: [number] representing index in references list

Citations: 4
References: 4

- "Personalized Spiking Neural Network Models of Clinical and Environmental Factors to Predict Stroke"

The paper [Dob+22] proposes a new method for the identification of associations between clinical and environmental time series: spiking neural networks. This paper is relevant to the present study, as it uses machine learning methods and also takes into consideration individual and family related factors.

Structure:

- Abstract - background, purpose of the study, details about used data and methods, results
- Introduction - medical and technical explanations, existing work and results, new approaches in this paper

- Methods - description of method, notations, schemas,
  - Method and System for Personalized Predictive Modeling on Integrated Personal Clinical Data and Dynamic Data of Environmental Changes - in detail explanation of the method, charts for data visualisation,
  - Encoding of Environmental Time-Series Data - description of encoding method
  - Environmental Data Mapping into a Personalized SNN Model - description of data mapping, variables and dimensions
  - Unsupervised Learning in the PSNN Model - method description
  - Supervised Learning, Classification and Prediction - method description
- Study Population and Datasets - details about involved data (quantitative, temporal, spatial, classifications by age, gender)
- Results - explained results, charts, figures
- Personalized Profiling of Individual Risk of Stroke Using Environmental Data
- Discussion - compares the obtained results to other studies, presents the advancement made by the study
- Conclusion - overview, future work
  - Acknowledgements
  - Author Contribution
  - Funding
- Declarations - legal aspects
  - Ethics Approval
  - Consent to Participate
  - Conflict of Interest
  - Open Access
- References - numbered list of cited and related work

In this paper, references appear as a list in the last section, in the following format: authors, title, journal, volume, no, pages, year. When cited, the format is: [number] representing index in references list

2.3 Original Contribution

The present study adds value to the research in the medical and meteorological fields as it gives an answer to an important question: Can stroke incidence be predicted using weather conditions? The answer to this question would help healthcare professionals save more patients hospitalized with this health emergency, contributing to the decrease of mortality due to stroke. Especially in Romania, where the study is conducted, this would have a great impact on the healthcare system overall. In comparison to existing work, the study uses artificial intelligence: neural networks and clustering algorithms to find the specified correlation, is focused on a small area, Transylvania, Romania and uses data collected over nine years 2013-2021.

3 Datasets

3.1 Meteorological Data

Numerical data is downloaded from Meteomanz.com and is collected from meteorological stations in Transylvania. The data contains collected values by days and by hours between Jan 2013 - Dec 2021. Data preprocessing includes formatting the values to be only numerical, removing redundant data. Because stroke incidence seems to be more related to the sudden differences in temperature and atmospheric pressure rather than their absolute values, these differences were computed using data collected by hours and added to the data by days dataset as columns for each line (day). The differences were computed for every 3 hours. For this process, the Pandas library from python was used.

3.2 Medical Data

The medical dataset was collected by healthcare professionals in Cluj-Napoca, Cluj, Romania between 2013-2021 and contains annonymous records of patients admitted with strokes. These records present data about the exact time of the stroke, the place where the patient was located at that time, generic data about the patient: gender, age, underlying health conditions.

4 Methods

4.1 Meteorological Data Visualization

To visualize the preprocessed meteorological data, the following libraries were used: Bokeh, Pandas, Matplotlib. The figures presented below represent the average values for temperature and atmospheric pressure over the studied time period and the 3 hours differences in these parameters.

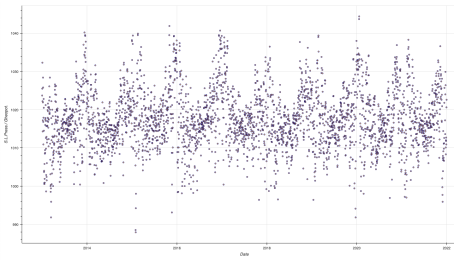

Figure 1: Average Atmospheric Pressure

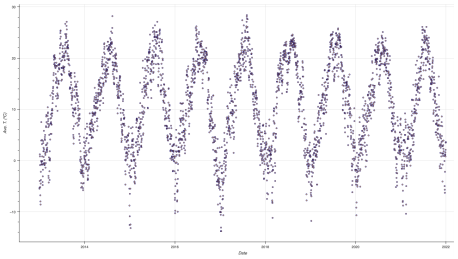

Figure 2: Average Temperature

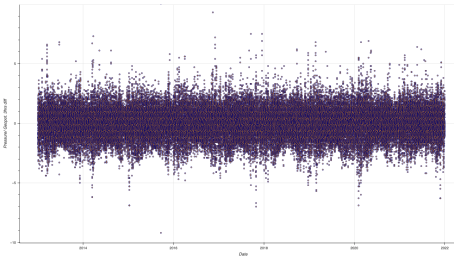

Figure 3: Atmospheric Pressure 3 hours Differences

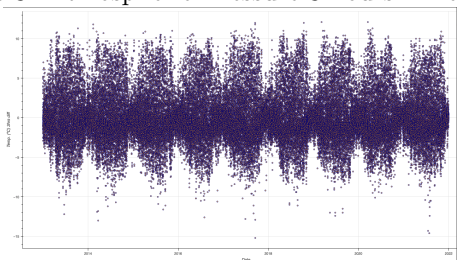

Figure 4: Temperature 3 hours Differences

4.2 Medical data Clustering

For grouping the data points representing patients, two clustering algorithms are used: K-Means Clustering and Mean-Shift Clustering. These unsupervised learning algorithms are used to identify groups of patients with similar features.

4.2.1 K-Means Clustering

4.2.2 Mean-Shift Clustering

4.3 Identifying Variations

Changes in the weather conditions are often times linked to atmospheric fronts passing over the area. Variations in weather parameters such as temperature and atmoshperic pressure can signal the passing of an atmoshperic front. These fronts can be classified into: cold front, warm front, stationary front and occluded front. Strokes incidence seems to be affected by variations in weather conditions, so identifying the moment a front passes over an area could lead to a prediction for stroke incidence. To verify this hypothesis, the maximum variations need to be first identified from the meteorological dataset. Two algorithms are used for this process: Feature Importance and Outliers Identification.

4.3.1 Feature Importance

4.3.2 Outliers Identification

4.4 Neural Network

The problem of predicting a sudden increase in the incidence of strokes can be seen as a classification problem and a possible approach for this is using an artificial neural network. This ANN receives as input the weather conditions for a day and outputs 1 or 0, 1 if this day corresponds to an increase in stroke incidence, i.e. the conditions for this day match the previously learned model for days correlated to a larger number of patients admitted, 0 otherwise.

5 Experimental Modelling

5.1 Data

To measure the performance of the algorithms and the validity of the hypothesis, multiple experiments are performed. The data is split into train, validate and test data. The train and validate data are used during the training phase of the algorithm, while the test data is used after the model is trained, to assess the obtained performance. The experiments are done using the validation data. Test data is used after the model is trained and the performance is assessed to manually test the algorithm, simulating its usage by an end user.

Data is collected as described in 3. As medical data contains records of strokes that occurred in multiple cities of Romania, multiple weather stations must be taken into consideration during the experiments. Not all cities appearing in the medical records have a weather station, so the closest weather station will be taken into consideration.

5.2 Metrics

Although the model needs to perform well, as it would be used in medical facilities and could affect the performance of the medical system, a small type I error is accepted, i.e. the model could predict a spike in the incidence of stroke, but in reality there is no spike. In this case, the medical staff would be prepared for stroke emergencies, but this does not have a negative impact on their performance. A type II error would mean that the model predicted no increase, so the doctors would not prepare for a spike that does appear in reality. This type of error should be minimized as much as possible in the model.

The following metrics are used for evaluating the experiments:

- Accuracy - number of correct predictions over all predictions
- Precision - how many positive predictions are true positives
- Recall - how many of the positives were detected as positive
- F1-Score - harmonic mean of precision and recall

Because the recall and precision should be balanced, the F1-Score is the most meaningful for the purpose of this study. The recall should be maximized, but without dropping the precision to a small value.

Value references that would validate the model:

- Accuracy > 90%
- Precision > 77%
- Recall > 85%
- F1-Score > 80%

5.3 Formal Model

Mathematically, the experiments require in the first step calculations of differences by hours of the following parameters, seen as variables: temperature (T), atmospheric pressure (AP), humidity (H), wind direction (WD), wind speed (WS). For this first computation the values for these parameters for each hour are used and 24 differences are computed for every day as follows:

For hour taking values from 0 to 24 and V being each one of the parameters (T, AP, H, WD, WS):

$$\begin{cases} E = (hour + 3)\%24 \\ S = hour \\ V\_S - E = V\_E - V\_S \end{cases} \tag{1}$$

The Artificial Neural Network used in the experiment computes the loss using the Cross Entropy loss function, where $\hat{y}$ is the probability of the output y = 1:

$$H(p, q) \;=\; -\sum_i p_i \log q_i \;=\; -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \tag{2}$$

The activation function used for the linear layers is ReLU:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

6 Case study

To illustrate the methodology and approach of the presented paper, a case study is performed using an initial, smaller dataset. For this, the medical and meteorological records from Cluj-Napoca city are taken into consideration. The experiment can then be extended to cover a larger area.

6.1 Data processing

Meteorological data is downloaded from Meteomanz.com, using the station code 15120, corresponding to Cluj-Napoca. The data covers the time interval 01.2013-12.2021. Medical data is collected from the public hospital in Cluj-Napoca and contains records from the same time interval.

The data is processed using the Pandas python library. More processing is needed for meteorological data, as separate files are downloaded for data by days and data by hours. This information is aggregated to compute the 3 hours differences used further. The parameters taken into consideration are: temperature (T), atmospheric pressure (AP), humidity (H), wind direction (WD), wind speed (WS). The medical records contain information about the day of the stroke occurrence, gender, age and medical history. Only 20% of the records give information about the exact time of the stroke, so this information is not kept, as completing the rest of the records with generated data would only induce errors. The data is split into 2 categories: train and test. The training dataset contains records from 07.2013-12.2021 and the testing one from 01.2013-06.2014.

6.2 Data visualization

6.2.1 Meteorological Data

Before applying the algorithm, data visualization was used to observe the days considered more probable to match with stroke occurrences.

In the first trial we consider the temperature and pressure differences for every 3 hours as the main meteorological parameters and draw a 3D plot with these values over the full time interval. As the data in this plot (5) is not easily readable, a 2D plot with color gradient is also made (6).

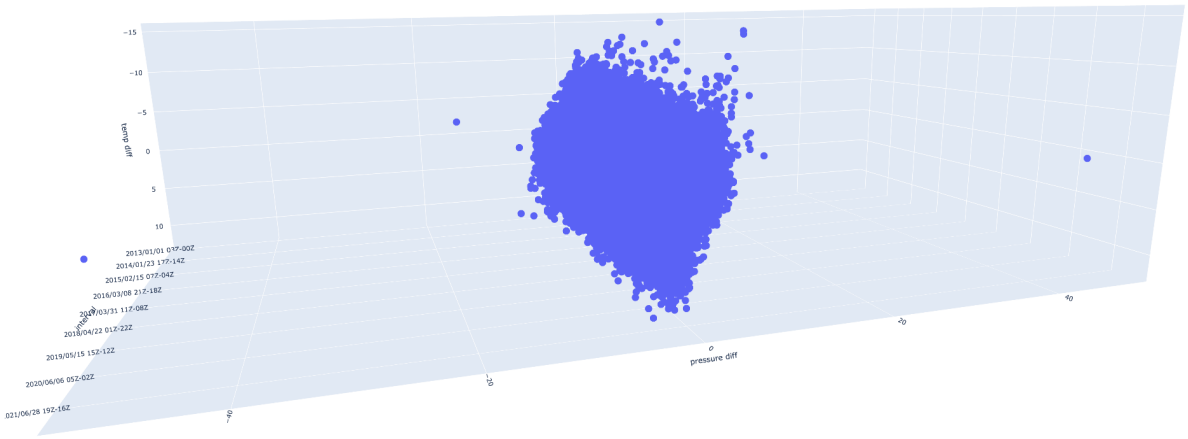

Figure 5: 3D plot of meteorological data

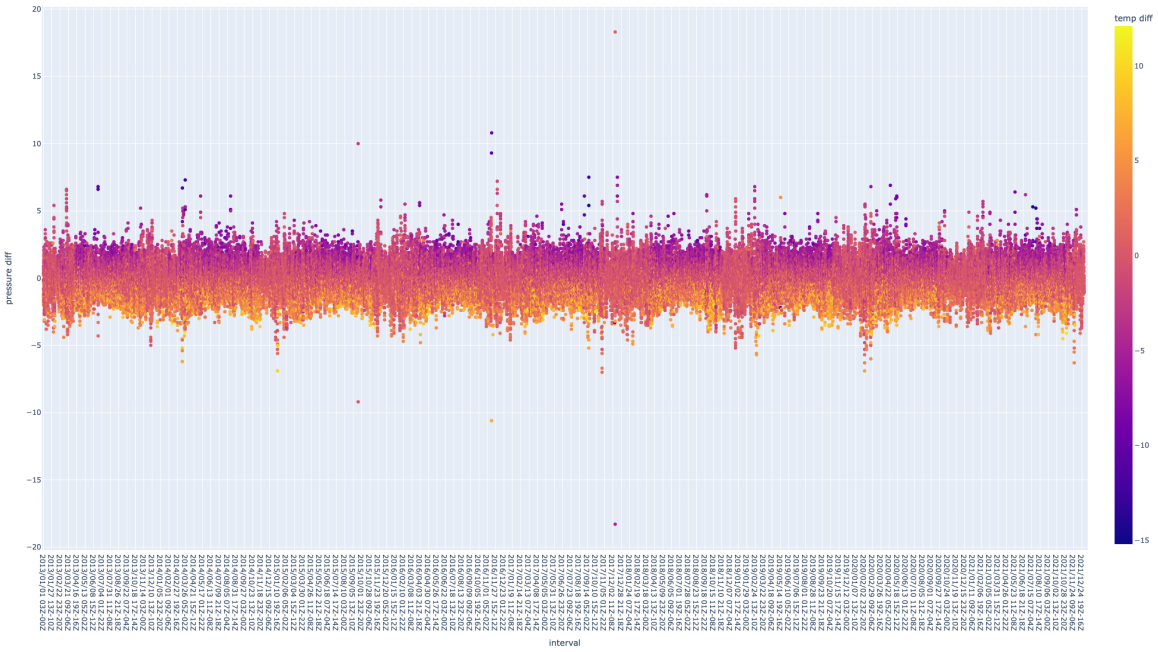

Figure 6: 2D plot of meteorological data

The visual result matches the expected one, some days across the time interval have greater temperature and pressure differences and a possible negative relationship between these two parameters can be observed. Most points with a larger, positive difference in pressure have a negative difference in temperature and the other way around. Most of the point are located in the [-2.5, 2.5] interval on the pressure diff axis, so these values will be considered normal. A few points are located above 10 or below -10, these values will be omitted as there is a high probability they are errors. The values in the intervals [-10, -2.5] and [2.5, 10], represented in Figures 7 and 8 will be analysed further.

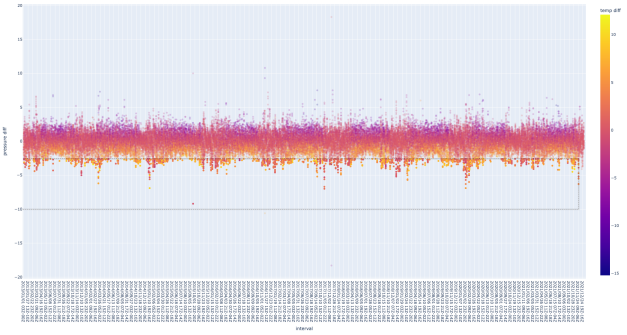

Figure 7: Interval [-10, -2.5]

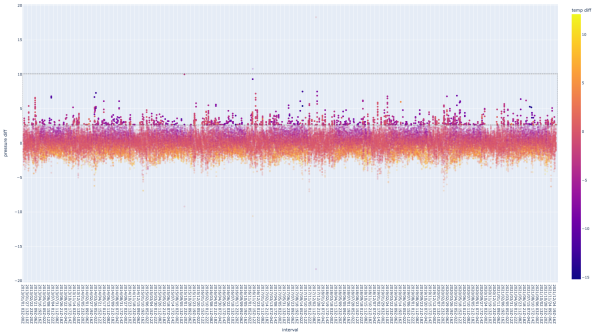

Figure 8: Interval [2.5, 10]

As it can be observed, some of these points have an extreme color for temperature difference as well, so these will be considered critical intervals. Days containing at least one critical interval are considered to be critical days.

6.3 Data labeling

The data is automatically labeled with the following approach: every day that has at least one stroke incidence is considered to be critical. The data is loaded from the medical and meteorological datasets and the labeling is done by searching for each day of the 9 years in the medical records. If found, the received label has the value 1, otherwise 0.

6.4 Experiment and Results

An Artificial Neural Network is used to classify days in one of the two classes: critical or non-critical. The neural network contains 7 Linear layers with ReLU being used as activation function for each one. The first layer has 120 input channels representing the 120 features of each day (24 differences for each of the 5 parameters) and the last layer has 2 output channels representing the 2 classes: critical and non-critical.

The model is trained with 15 epochs and the datasets are split and shuffled into batches of size 32. This increases the probability for the critical days to be uniformly distributed across batches.

For each epoch, the train accuracy(11), train loss (10) and test accuracy are computed and plotted. If the current value for the test accuracy is greater than the best one until this iteration, the model is saved as the best one (9). During the validation phase, a confusion matrix is computed and the one for the best model is also plotted after the training is done. To better visualize the accuracy by classes and observe possible overfitting to one class, the evolution of these values by epochs are also plotted (12).

Results show a high accuracy, but, as explained in 5.2, the False Negative values have a high impact on the performance of the model and the Recall is more relevant than the accuracy. In this experiment, the recall is 0, as there is no True Positive prediction. This behavior could appear as a result of a very large dataset with very few examples for positive classification. This is a possible explanation for our experiment as there are less than 70 positive values in more than 3000 records. An improvement could be made by simplifying the input data. The current input has 120 channels containing data about 5 weather parameters in 24 3 hours differences

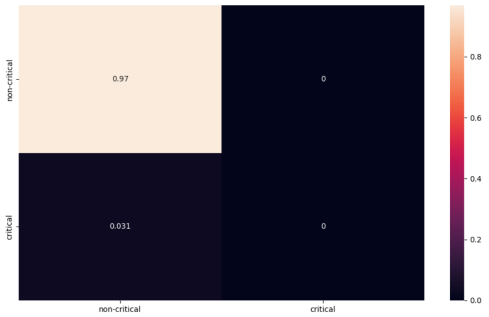

Figure 9: Confusion Matrix

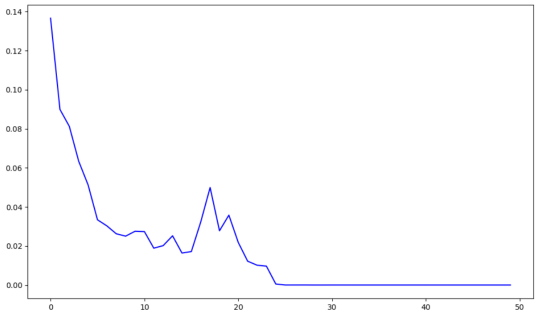

Figure 10: Train Loss

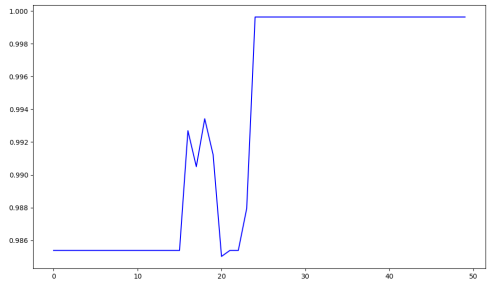

Figure 11: Train Accuracy

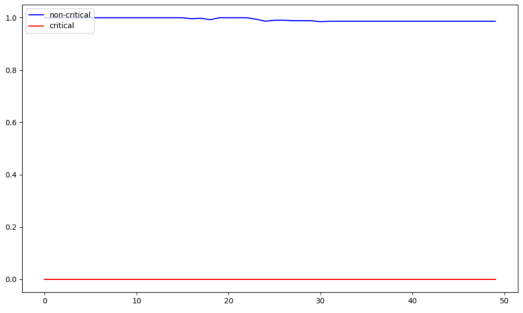

Figure 12: Accuracy By Classes

for each day. The input could contain only 5 input channels, one per weather parameter, and its value can be computed as the maximum difference for that parameter during the day. This approach could make a critical day more easily identifiable.

7 Related Work - Comparison

7.1 Real Dataset

In order to make a relevant comparison between our paper and related work in this field, the dataset used in the experiments must be extended. For this, more meteorological data is collected from multiple weather stations across Transylvania and more medical records are taken into consideration, with the place of the incident being outside of Cluj-Napoca. To put all this data together, every stroke incident must be associated with the closest weather station that provides data for the specific day of the incident. This adds a new processing step that was not needed for the small data set covering Cluj-Napoca only.

7.2 Approaches and Metrics

There are two main approaches for the presented problem in existing literature: statistical calculus and artificial intelligence, most studies using the first one. The data used in each paper is different, as it strongly depends on the geographical area covered by the study. Thus, no comparison can be made between our study and related work on the same data, but overall performance and approaches are relevant even if the input data is different.

One of the most relevant studies on stroke incidence correlated to weather conditions is by Knezovic et al. The paper is focused on data collected from Zagreb, Croatia, which has a similar climate to Romania, the area observed for the present study. However, the approach in [Kne+18] is very different, focusing on statistical methods, with no artificial intelligence involved. Their results show a seasonal variation of stroke occurrence, different per stroke subtype.

A study with an approach more similar to ours is [Li+14]. The approach in this paper is centered around a Back Propagation Neural Network (BPNN), so although the used data is different, the methodology is similar, so the results are relevant. Fang et al. obtained a relative error of 6% for 2010 early January, when the real value was 3.2 and the prediction was 3, a relative error of 7% for mid January 2010, when the predicted incidence rate was 2.7 and the real one was 2.9. Overall, their predictions had an error between 6% - 8% and is considered "accurate and reasonable"([Li+14]).

8 Ethics

No medical data used in the present paper was collected without the consent of the patients and all data is anonymous.

9 Discussion

10 Conclusion

References

[Ber+89] V M Berginer et al. “Clustering of strokes in association with meteorologic factors in the Negev Desert of Israel: 1981-1983”. In: *Stroke* 20 (1989), pp. 65–69. DOI: 10.1161/01.STR.20.1.65.

[FH02] Thalia S. Field and Michael D. Hill. “Weather, Chinook, and Stroke Occurrence”. In: *Stroke* 33 (2002), pp. 1751–175. DOI: 10.1161/01.STR.0000020384.92499.59.

[WSK02] Hongbing Wang, Michikazu Sekine, and Xiaoli ChenSadanobu Kagamimori. “A study of weekly and seasonal variation of stroke onset”. In: *Int J Biometeorol* 47 (2002), pp. 13–20. DOI: 10.1007/s00484-002-0147-x.

[Jim+08] J. Jimenez-Condea et al. “Weather as a Trigger of Stroke”. In: *Cerebrovasc Dis* 26 (2008), pp. 348–354. DOI: 10.1159/000151637.

[R+11] Magalhães R et al. “Are Stroke Occurrence and Outcome Related to Weather Parameters? Results from a Population-Based Study in Northern Portugal”. In: *Cerebrovasc Dis* 32 (2011), pp. 542–551. DOI: 10.1159/000331473.

[Hor+12] Aya Hori et al. “Effects of weather variability and air pollutants on emergency admissions for cardiovascular and cerebrovascular diseases”. In: *International Journal of Environmental Health Research* 22 (2012), pp. 416–430. DOI: 10.1080/09603123.2011.650155.

[Kas+14] Nikola Kasabov et al. “Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke”. In: *Neurocomputing* 134 (2014), pp. 269–279. DOI: 10.1016/j.neucom.2013.09.049.

[Li+14] Fang Li et al. “An Improved Back Propagation Neural Network Model and Its Application”. In: *JOURNAL OF COMPUTERS* 9(8) (2014). DOI: 10.4304/jcp.9.8.1858-1862.

[Cev+15] Yunsur Cevik et al. “The association between weather conditions and stroke admissions in Turkey”. In: *International Journal of Biometeorology* 59 (2015), pp. 899–905. DOI: 10.1007/s00484-014-0890-9.

[AD+17] Tarnoki AD et al. “Relationship between weather conditions and admissions for ischemic stroke and subarachnoid hemorrhage”. In: *Croat Med J* 58 (2017), pp. 56–62. DOI: 10.3325/cmj.2017.58.56.

[Don18] Eric S Donkor. “Stroke in the 21st Century: A Snapshot of the Burden, Epidemiology, and Quality of Life”. In: *Stroke Res Treat.* 27 (2018). DOI: 10.1155/2018/3238165.

[Kne+18] Marijana Knezovic et al. “The role of weather conditions and normal level of air pollution in appearance of stroke in the region of Southeast Europe”. In: *Acta Neurologica Belgica* 118 (2018), pp. 267–275. DOI: 10.1007/s13760-018-0885-0.

[N+20] Matsumaru N et al. “Weather Fluctuations May Have an Impact on Stroke Occurrence in a Society: A Population-Based Cohort Study”. In: *Cerebrovasc Dis Extra* 10 (2020), pp. 1–10. DOI: 10.1159/000505122.

[A+21] Ponmalar A et al. “Stroke Prediction System Using Artificial Neural Network”. In: *International Conference on Communication and Electronics Systems* 6 (2021), pp. 1898–1902. DOI: 10.1109/ICCES51350.2021.9489055.

[Dob+22] Maryam Doborjeh et al. “Personalized Spiking Neural Network Models of Clinical and Environmental Factors to Predict Stroke”. In: *Cogn Comput* (2022). DOI: https://doi.org/10.1007/s12559-021-09975-x.

[Nie+22] Stefan Niebler et al. “Automated detection and classification of synoptic-scale fronts from atmospheric data grids”. In: *Weather Clim. Dynam* 3 (2022), pp. 113–137. DOI: 10.5194/wcd-3-113-2022.