Comparative Analysis of Mice Protein Expression Data: Assessing Genotype and Behavioral Treatments Using Machine Learning Algorithms

Surendiran Balasubramanian[1(✉)], Malarupu Charan Sai[2], M. Dheeraj Kumar[2], Kunuthuru Karthik Kumar Reddy[2], Veera Harish Muthazhagu[1], and Ramanathan Palaniappan[2]

[1] National Institute of Technology Puducherry, Karaikal 609609, Puducherry, India
surendiran@nitpy.ac.in

[2] Madanapalle Institute of Technology and Science, Madanapalle 517325, Andhra Pradesh, India

Abstract. The study uses machine learning methods to examine the effects of genotype and behavioral interventions on the patterns of protein expression in mice. The dataset includes measurements of the expressions levels of 77 proteins in the cerebral cortex of 72 mice, comprising 34 trisomic (Down syndrome) mice and 38 normal mice. A total of 1080 measurements were made for each protein, with 15 measurement per sample. Based on genotype, behavior, and treatment, the mice are divided into eight different groups: trisomy mice, control mice stimulated to learn and injected with saline (c-CS-s), control mice stimulated to learn and injected with memantine (c-CS-m), and control mice not stimulated to learn and injected with saline (c-SC-s). Memantine was injected into control mice that had not been stimulated to learn (c-SC-m), memantine was injected into trisomy mice that had been stimulated to learn and given saline (t-CS-s), memantine was injected into trisomy mice that had been stimulated to learn and given memantine (t-CS-m), and trisomy mice had not been stimulated to learn and given saline (t-SC-s). In this work, four distinct classification algorithms-Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting-were used to sort the mice into their appropriate classifications. The goal was to evaluate each algorithm's capability and accuracy in assigning precise class labels using protein expression data. We assessed the effectiveness and accuracy of different algorithms in completing this categorization assignment through a comparative study.

Keywords: Mice protein expression · Genotype · Behavioral treatments · Machine learning algorithms · Cognitive impairment

P. Pareek et al. (Eds.): IC4S 2024, LNICST 597, pp. 257–267, 2025.
https://doi.org/10.1007/978-3-031-77075-3_21

1 First Introduction

People with Down syndrome have cognitive and learning disabilities as a result of a hereditary disease. Researchers are trying to figure out why this occurs and how they can support people with Down syndrome. Mice are studied because their genes are comparable to those of humans. In our investigation, we examined a unique dataset that provides information on the concentrations of 77 proteins in mouse brains. Some mice and others exhibit Down syndrome. Additionally, some mice had been educated to pick up new skills, while others hadn't. Based on these factors, we divided all of the mice into various groups. For our research, we had two main objectives. First, depending on the protein levels, we employed clever computer programs (machine learning techniques) [1] to determine which group each mouse belonged to. This made it easier for us to grasp how the features of the mice are related to protein patterns. To identify any natural patterns or groups in the data without using the group labels, we employed another technique called clustering. This helped us comprehend how the mice differ or are similar to one another. Our findings may assist in identifying significant cognitive issue markers and shed light on how various treatments alter the brain proteins of mice. Future research might be steered by this information, which could also serve to advance care for Down syndrome patients. The data we used and the analysis process will be covered in more detail in the following sections of our study. We will also talk about our results and what they suggest for the future.

2 Literature Survey

Analysis of protein expression data has become a crucial area in biomedical research, allowing for the clarification of complex biological mechanisms. Sri et al.'s [2] work illustrated the value of categorization based on protein expression levels in identifying proteins essential to trisomic mice's capacity for learning. This research served as a starting point for our inquiry into related classification methods. In a comparative examination of mice protein expression, Witten et al. [3] used both clustering and classification techniques. Their research played a crucial role in our decision to investigate a variety of algorithms [4], which resulted in the effective application of Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting Classification approaches.

Furthermore, Higuera et al. [5] used self-organizing feature maps to discover proteins essential for learning in a mouse model of Down syndrome, which paved the path for our data preprocessing techniques. In order to improve the integrity of our dataset, we borrowed their method for dealing with missing values. Alickovic et al.'s [6] study of data mining methods for classifying medical data focused on the use of Support Vector Machines. We added Support Vector Machine techniques and, using their insights as a guide, obtained remarkable accuracy, demonstrating the dependability of our methods.

We were inspired by the larger field of neural networks as we expanded our research into the area of deep learning. Recurrent neural networks (RNNs) are

increasingly being used in a variety of data processing situations, where temporal dependencies and sequential patterns are important considerations [2]. This innovative addition deepened our analysis and enabled us to identify latent patterns in the complicated protein expression data.

The literature review as a whole emphasizes the value of categorizing protein expression data [7] in the context of biological and biomedical research. By examining categorization strategies, data preprocessing approaches, and the possibility of deep learning, existing studies have established a solid basis. Our study expands on previous contributions by combining conventional algorithms with cutting-edge methodologies to offer a thorough examination of the complex connections between a mouse's genotype, behavior, and protein expression patterns [8].

3 Methodology

3.1 Dataset Acquisition

The "Mice Protein Expression" dataset that was obtained from Kaggle served as the foundation for this investigation. The expression levels of 77 proteins make up the dataset. There are a total of 72 mice, comprising 34 trisomic (Down syndrome) mice and 38 control mice. Figure 1 shows the genotype classification of mice based on treatment and behavioral patterns. Throughout the studies, 15 measurements of each protein per mouse were taken. The collection contains 1080 measurements for each protein. The Kaggle availability of the dataset ensures its availability and dependability for research purposes [5].

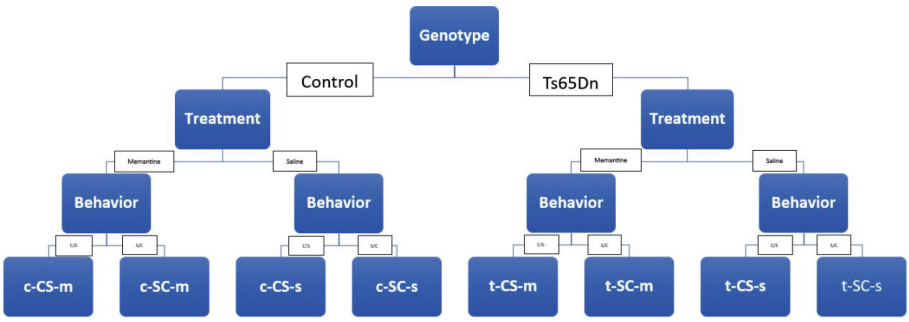

Fig. 1. Genotype classification of mice based on treatment and behavioral patterns

3.2 Data Preprocessing

To ensure that the dataset was suitable for the chosen models and of high data quality, it underwent intensive preparation procedures. First, irrelevant columns

that added little to nothing to the goal of classifying students were removed, which decreased error and improved accuracy. Standardization (StandardScaler) and normalization (min-max scaling) are two popular scaling techniques that allow fair comparisons between various properties (features) as well as converting the data into a format that the algorithms can use.

3.3 Data Split

An 80/20 split was used to divide the preprocessed dataset into training and testing sets. The training set received 80% of the data with this split ratio, while the testing set only received 20%. The machine learning models can be trained using the more varied samples from the larger training set. Unobserved data from the testing set was used to evaluate the models' performance as well as their capacity to generalize beyond the training set. Machine learning studies frequently use the 80/20 split, which strikes a compromise between offering sufficient training data and guaranteeing a correct judgment on circumstances that have not yet been observed. Figure 2 shows the dataset statistics of each class.

Class	Training Set	Testing Set	Total
t-CS-m	108	27	135
t-SC-m	108	27	135
t-CS-s	84	21	105
t-SC-s	108	27	135
c-CS-m	120	30	150
c-SC-m	120	30	150
c-CS-s	108	27	135
c-SC-s	108	27	135

Fig. 2. Dataset Statistics of each class

3.4 Feature Selection

In machine learning and deep learning, feature selection is a crucial stage that may have an impact on the models' accuracy. Feature selections are provided in two different ways for each algorithm, such as Decision Tree, Random Forest, Support vector machine (SVM) [9], Gradient Boosting Machine (GBM), and Recurrent Neural Network (RNN), which is used to classify mice in the dataset. Only 77 characteristics are provided for first models. The dataset now has two additional features, genotype and therapy. The genotype characteristic identifies a mouse as either trisomic (t) or control (c). The treatment attribute specifies whether the mouse received memantine (m) or saline (s) treatment. These two characteristics, in addition to the 77(proteins) characteristic, are used to group the mice into one of eight groups:t-CS-s, t-CS-m, t-SC-s, t-SC-m, and discovered the comparative accuracy of each model using two techniques of selecting features.

3.5 Model Training

Training data refers to the act of instructing a machine learning algorithm to make predictions or decisions based on a specific set of data. To determine a mouse's class based on its protein expression levels (First) and other factors like genotype and treatment1 (Second), machine learning algorithms are trained using the Kaggle Mice Protein Expression dataset. The dataset is used to produce two sets: a training set and a test set. The algorithm is tested on the test set after having been trained on the training set. In order to train a model, you must give the algorithm the necessary input characteristics (such as the genotype, treatment, and levels of protein expression) as well as the relevant output (the mouse class) [10]. The algorithm uses this information to generate predictions based on the input features.

3.6 Model Evaluation

On our protein expression dataset, we thoroughly assessed five different algorithms: Decision Tree, Random Forest, Support Vector Machine (SVM), Gradient Boosting, and Recurrent Neural Network (RNN) [11]. Every algorithm was rigorously tested, yielding observable accuracy rates and showcasing their potential for categorizing mice based on protein expression and contextual cues.

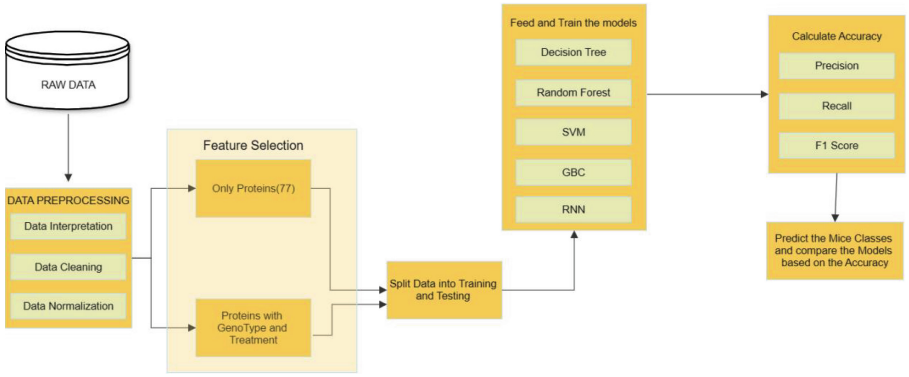

Fig. 3. Mice Protein expression classification process flow

3.7 Comparative Analysis

The standard Decision Tree, while successful, gains greatly from the addition of genotype and treatment variables, increasing its accuracy from 0.8739 to 0.9640, as we saw when comparing the algorithms. Figure 3 shows the mice protein expression classification process flow. An ensemble of decision trees called Random Forest regularly produced accuracy results up to 0.9989, demonstrating

robustness in handling large datasets. In high-dimensional data categorization, SVM and Gradient Boosting demonstrated outstanding accuracy rates of up to 0.9993 and 1.00, respectively. By successfully capturing sequential patterns and reaching accuracy rates of 0.5586 and 0.9910 with and without contextual characteristics, respectively, the incorporation of RNN brought a distinctive temporal dimension.

4 Machine Learning Algorithms

Four machine learning algorithms: Decision Tree, Random Forest, SVM, and Gradient Boosting Classification were originally trained and assessed. We used a ten-fold method of feature selection, taking into account both the level of protein expression alone and additional characteristics like genotype and therapy. The datasets were divided into training and testing sets, with the training data being used to train the algorithms. On the basis of the testing data, the developed models were evaluated, and several performance measures, including as accuracy, precision, and recall, were computed [12]. Notably, adding genotype and treatment information enhanced all algorithms' performance, proving their value in raising classification accuracy.

4.1 Deep Learning with Recurrent Neural Network (RNN)

The dataset was preprocessed to suit the input requirements of the Recurrent Neural Network (RNN), reshaping the data into sequences. The RNN model consisted of a SimpleRNN layer followed by densely connected layers. The model was compiled with appropriate loss functions and optimizers, and trained on the preprocessed data. The RNN demonstrated promising results, showcasing its potential for capability.

These algorithms' comparison analyses revealed vital knowledge on their effectiveness in identifying protein expression patterns associated with cognitive qualities in mice, revealing their individual strengths and shortcomings. The next part provides a thorough assessment of the suggested approaches by elaborating on the experimental findings. Keywords: Deep Learning, Recurrent Neural Network, Gradient Boosting Classification, Decision Tree, Random Forest, Support Vector Machine, Genotype, Cognitive Abilities, and Treatment.

5 Deep Learning Architecture

After thorough data preprocessing, which included cleaning missing values and encoding categorical features, we explored various algorithms and implemented Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting Classification. Additionally, we introduced a novel aspect by applying a Recurrent Neural Network (RNN) model to the protein expression data in mice. Our results contributed to a better comprehension of the complex interactions

between genotype, behavior, and protein expression patterns in mice by demonstrating the superiority of some algorithms and the promise of RNNs in this setting.

We did an in-depth examination of classifying protein expression data in mice, adopting a multidimensional strategy that encompassed multiple machine learning approaches. Our investigation begins with thorough data preprocessing, comprising the cleaning of missing values and the encoding of categorical variables like 'Genotype' and 'Treatment'. Then, using a variety of methods, including Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting Classification, we started a thorough investigation. Each algorithm was painstakingly adjusted and tested to determine how well it classified the complex patterns of protein expression found in the mice dataset. The development of a cutting-edge Recurrent Neural Network (RNN) model, a rapidly developing field in deep learning, is what really distinguishes our research. This brand-new architectural paradigm gave our study a cutting-edge component. The RNN architecture that we used had two layers: a SimpleRNN layer with 64 units and a Dense layer with 32 units after that. The ultimate layer, characterized by softmax activation, proficiently predicted the class labels, thereby facilitating an enhanced comprehension of the underlying patterns and relationships within the data. This hybridized approach, combining traditional machine learning algorithms with the advanced capabilities of RNNs, yielded a comprehensive framework for assessing and comparing classification performance. Our rigorous research not only reinforced the superiority of specific algorithms in capturing the complexity of the data but also highlighted the enormous potential of RNNs in understanding the intricate interplay between genotype, behavior, and protein expression patterns in mice. Our discoveries have substantial implications for several fields, including biology, genetics, and machine learning, highlighting the broad effect of our research project. Our research provides a comprehensive view of the opportunities and challenges in categorizing complicated biological data by integrating conventional methods with cutting-edge paradigms.

5.1 Unveiling Patterns with Recurrent Neural Network

In order to further our analysis, we implemented a Recurrent Neural Network (RNN) into the classification framework. Surprisingly, the RNN demonstrated distinct capabilities in capturing sequential dependencies within the data, achieving an accuracy of 0.5586 without context, which notably increased to 0.9910 upon inclusion of genotype and treatment data.

6 Result and Analysis

6.1 Performance of Decision Tree and Random Forest

The dataset was initially classified using the Decision Tree technique, with an accuracy of 0.8739. The accuracy dramatically increased to 0.9640 after incorporating genotype and treatment data, demonstrating the value of contextual

features in improving classification results. The Random Forest algorithm next proved its prowess by producing impressive accuracies for the two cases of 0.9905 and 0.9989. Decision trees' collective knowledge was harnessed by Random Forest's ensemble structure, improving its predictive power. Table 1 shows the comparative analysis of models only on Proteins feature and Table 2 shows comparative analysis of models on Proteins with Genotype and Treatment feature.

6.2 Effectiveness of Support Vector Machine and Gradient Boosting

The Support Vector Machine (SVM) and Gradient Boosting approaches, both of which exhibit remarkable performance, were also included in our investigation. Gradient boosting increased accuracy to 0.9819 and 1.00, whereas SVM attained accuracy of 0.9945 and 0.9993. The inclusion of genotype and treatment attributes boosted classification accuracy in a consistent manner, confirming the importance of these contextual features in identifying complex patterns in the data.

Table 1. Comparative analysis of models only on Proteins feature

Models	accuracy	precision	recall	F1score
Decision Tree	0.873	0.886	0.874	0.874
Random Forest	0.990	0.989	0.990	0.989
SVM	0.982	0.983	0.982	0.982
GBC	0.981	0.982	0.981	0.981

Table 2. Comparative analysis of models on Proteins with Genotype and Treatment features

Models	accuracy	precision	recall	F1score
Decision Tree	0.964	0.970	0.964	0.964
Random Forest	0.989	0.998	0.999	0.998
SVM	1.00	1.00	1.00	1.00
GBC	1.00	1.00	1.00	1.00

6.3 Comparative Insights and Future Prospects

When comparing the algorithms, Random Forest and SVM stand out as strong competitors, especially in contexts with more information. The robustness of the Gradient Boosting technique is demonstrated by its great accuracy performance. The RNN offers a distinctive viewpoint that further supports the advantages of incorporating contextual features thanks to its ability for sequential learning. Figure 4 shows the comparison of each model over accuracy.

6.4 Implications and Future Directions

The results of the experiment have important ramifications for comprehending biological events. The capacity to precisely categorize mice based on contextual factors and protein expression may one day be extended to the early detection of learning impairments. Additionally, the effective integration of deep learning algorithms provides avenues for additional investigation in the interpretation of biological data, pushing the limits of our knowledge and prospective applications.

6.5 Analysis

Decision Tree vs. Random Forest. In all feature sets, the Random Forest method fared better than the Decision Tree approach, achieving greater accuracy, precision, recall, and F1-score values. To enhance classification performance and manage overfitting, Random Forest uses numerous decision trees, leading to greater generalization.

Genotype and Therapy are Important. Including genotype and therapy as extra parameters greatly increased classification accuracy across all algorithms. The models' improved ability to distinguish between various genotype and treatment categories resulted in improved performance.

SVM and Gradient Boosting. When using the entire feature set (proteins + genotype + treatment), SVM and Gradient Boosting both obtained astounding accuracy of 100. These algorithms showed extraordinary capacity to recognize links and patterns between features, leading to accurate classification.

Robustness and Generalization. All algorithms performed well overall, proving their robustness in classifying mice based on protein expression data. Furthermore, adding domain-specific variables produced models with better generalization, fostering accurate categorization across many classes.

Clinical Importance. The study's results provide important new information on the crucial roles that protein expression, genotype, and treatment play in cognitive performance, particularly in Down syndrome mice. These understandings help to clarify the underlying mechanisms and may help to direct therapeutic procedures.

7 Discussion And Conclusion

7.1 Discussion

In this study, we investigated the use of various machine learning methods to categorize mice based on protein expression data. We found that using the Decision Tree algorithm with genotype and treatment considerably enhanced accuracy, reaching 96.40% compared to 87.39% with proteins just, by comparing

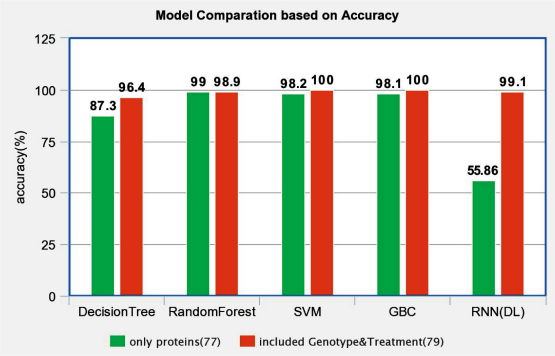

Fig. 4. Comparison of model over accuracy

our findings to earlier works [13]. Additionally, our results showed that Gradient Boosting Classification performed remarkably well, obtaining an astounding accuracy of about 100%. These results underline how crucial it is to use strong ensemble algorithms and incorporate pertinent information to improve classification accuracy. Our models provide useful insights for therapeutic interventions, and the discovered essential proteins, genotype, and treatment hold potential clinical value for understanding learning ability in Down syndrome mice. We do accept that the study's shortcomings, such as the quantity of the dataset and the nature of the experiments, call for more research. Overall, this study adds to the body of knowledge on how proteins are expressed and how they affect cognitive function. It also offers a potential method for further research in fields related to cognitive science.

7.2 Conclusion

Using different machine learning techniques, including Decision Tree, Random Forest, Support Vector Machine (SVM), and Gradient Boosting Classification, our study examined the classification of mice based on protein expression data. We examined the impact of various feature sets on classification performance, building on past research [14]. The outcomes showed that including genotype, therapy, and protein expression together with gene expression considerably increased the accuracy of all algorithms, particularly obvious in Decision Tree and Gradient Boosting, which attained almost flawless accuracy. This finding has major implications for cognitive research and potential treatment therapies since it sheds light on crucial proteins and their function in learning ability in Down syndrome mice. The impact of our project lies in showcasing the potential of machine learning algorithms in enhancing the classification of complex biological data, paving the way for developments in the field of cognitive rehabilitation. Our work demonstrates the pivotal role of feature selection in optimizing classification accuracy, setting the stage for future studies to leverage this knowledge for improving understanding and treatment strategies for cognitive impairments

in genetic disorders. Overall, our study adds to the expanding body of knowledge in the field of cognitive research and emphasizes the ability of machine learning to extract key information from complex datasets, with implications for both fundamental science and clinical practice.

References

1. Salzberg, S.L.: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc. (1993). Mach Learn 16, pp. 235-240 (1994)
2. Sri, S., Preethi, S., Shamruthi, R., Shana, J.: Classification based on protein expression levels and identification of proteins critical to learning ability in Trisomic mice. In: Innovations in Power and Advanced Computing Technologies (i-PACT), pp. 01-07 (2021)
3. Witten, I.H., Frank, E., Hall, M.A.: Data mining: practical machine learning tools and techniques**5**, 1–2 (2006)
4. Nguyen, C., Costa, A., Cios, K., Gardiner, K.: machine learning methods predict locomotor response to MK-801 in mouse models of down syndrome. J. Neurogenet. **25**(1–2), 40–51 (2011)
5. Higuera, C., Gardiner, K., Cios, K.: Mice protein expression. UCI Machine Learning Repository (2015)
6. Alickovic, E., Subasi, A.: Data Mining Techniques for Medical Data Classification (2011)
7. Ribeiro-Machado, C., Silva, S.C., Aguiar, S., Faria, B.M.: Protein attributes-based predictive tool in a down syndrome mouse model: a machine learning approach. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) WorldCIST'18 2018. AISC, vol. 747, pp. 19–28. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-77700-9_3
8. Ahmed, M.M., et al.: Protein dynamics associated with failed and rescued learning in the Ts65Dn mouse model of Down syndrome. PLoS ONE **10**(3), e0119491. (2015)
9. Eicher, T., Sinha, K. : A support vector machine approach to identification of proteins relevant to learning in a mouse model of Down Syndrome. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 3391–3398 (2017)
10. Al-Rashid, S.Z.: Studying the effect of Mouse models for gene expression using coregionalization models in Gaussian process. In: Scientific International Conference Najaf (SICN), pp. 210-215 (2019)
11. Li. F., Zhou H.: Predicting mouse transmembrane protein types based on the increment of diversity combined with the support vector machine. In: 3rd International Conference on Biomedical Engineering and Informatics, pp. 2226–2229 (2010)
12. Hofmann, M., Klinkenberg, R.: RapidMiner: Data Mining Use Cases and Business Analytics Applications, 1st edn. CRC Press (2013)
13. Saringat, M.Z., Mustapha, A., Andeswari, R.: Comparative analysis of mice protein expression: clustering and classification approach. In: International Journal of Integrated Engineering, vol. 10 (2018)
14. Samsudin, N.A., Bradley, A.P.: Extended Naïve Bayes for group based classification. In: Herawan, T., Ghazali, R., Deris, M.M. (eds.) Recent Advances on Soft Computing and Data Mining. AISC, vol. 287, pp. 497–505. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07692-8_47