Published in final edited form as:

Proc (Int Conf Comput Sci Comput Intell). 2021 December ; 2021: 349–354. doi:10.1109/csci54926.2021.00130.

Application of Machine Learning to Sleep Stage Classification

Andrew Smith,
Department of Computer Science and Engineering (University of South Carolina), Columbia, SC 29208 USA

Hardik Anand,
Department of Computer Science and Engineering (University of South Carolina), Columbia, SC 29208 USA

Snezana Milosavljevic,
Department of Pharmacology, Physiology, and Neuroscience University of South Carolina School of Medicine, Columbia, SC 29208 USA

Katherine M. Rentschler,
Department of Pharmacology, Physiology, and Neuroscience University of South Carolina School of Medicine, Columbia, SC 29208 USA

Ana Pocivavsek,
Department of Pharmacology, Physiology, and Neuroscience University of South Carolina School of Medicine, Columbia, SC 29208 USA

Homayoun Valafar
Department of Computer Science and Engineering (University of South Carolina), Columbia, SC 29208 USA

Abstract

Sleep studies are imperative to recapitulate phenotypes associated with sleep loss and uncover mechanisms contributing to psychopathology. Most often, investigators manually classify the polysomnography into vigilance states, which is time-consuming, requires extensive training, and is prone to inter-scorer variability. While many works have successfully developed automated vigilance state classifiers based on multiple EEG channels, we aim to produce an automated and openaccess classifier that can reliably predict vigilance state based on a single cortical electroencephalogram (EEG) from rodents to minimize the disadvantages that accompany tethering small animals via wires to computer programs. Approximately 427 hours of continuously monitored EEG, electromyogram (EMG), and activity were labeled by a domain expert out of 571 hours of total data. Here we evaluate the performance of various machine learning techniques on classifying 10-second epochs into one of three discrete classes: paradoxical, slow-wave, or wake. Our investigations include Decision Trees, Random Forests, Naive Bayes Classifiers, Logistic Regression Classifiers, and Artificial Neural Networks. These methodologies have achieved accuracies ranging from approximately 74% to approximately 96%. Most notably, the Random Forest and the ANN achieved remarkable accuracies of 95.78% and 93.31%, respectively. Here we

aks3@email.sc.edu .

have shown the potential of various machine learning classifiers to automatically, accurately, and reliably classify vigilance states based on a single EEG reading and a single EMG reading.

Index Terms—

sleep-scoring; machine learning; artificial intelligence; neuroscience; electrophysiology

I. Introduction

Nearly 70 million Americans are afflicted by chronic sleep disorders or intermittent sleep disturbances that negatively impact health and substantially burden our health system.Sleep is essential for optimal health. Sleep is one of the most critical and ubiquitous biological processes, next to eating and drinking. It has been shown that there is no clear evidence of the existence of an animal species that does not sleep [1]. Sleep constitutes about 30% of the human lifespan. Assessment of sleep quality is multifactorial and is composed of adequate duration, good quality, appropriate timing and regularity, and the absence of sleep disturbances or disorders. Sleep duration is used as a metric to describe the standard of healthy sleep. The American Academy of Sleep Medicine (AASM) and Sleep Research Society (SRS) issued a consensus statement recommending “adults should sleep 7 or more hours per night on a regular basis to promote optimal health” [2]. However, sufficient sleep is severely undervalued as a necessary biological process for maintaining proper mental health. A recent survey from the Center for Disease Control and Prevention (CDC) found that only 65% of adults reported a healthy duration of sleep [3].

From a translational perspective, animal studies are imperative to recapitulate phenotypes associated with sleep loss (hyperarousal, cognitive impairment, slowed psychomotor vigilance, behavioral despair) and uncover mechanisms contributing to psychopathology, with the added benefit of homogeneity within rodent subjects [4], [5]. Sleep studies are readily conducted in small animals by implanting electrodes to obtain electroencephalogram (EEG) and electromyogram (EMG). Sleep is categorized into two major classes, non-rapid eye movement (NREM) and rapid eye movement (REM) sleep, and arousal is classified as wake. Most often, investigators manually classify the polysomnography into vigilance states, and this practice is time-consuming and also greatly limits the size of a study’s data set. To accurately classify vigilance states, investigators undergo extensive training, yet the subjective nature of classifying limits inter-scorer reliability.

Several automated vigilance state classifiers have been established, and nearly all of these algorithms rely on multi-channel EEG data and local field potential (LFP) signaling oscillations within the brain [6]–[12]. The advantages of multi-channel systems are outweighed by the disadvantage of tethering small animals to transmit signals via wired connections to computer programs. Tethered animals are combating confounds including limited mobility within recording cages and potential impacts on natural sleep states [13]–[15]. For these reasons, it is advantageous to automate vigilance state classification with telemetric battery devices that are surgically implanted to open a single EEG and EMG from each small animal. Consistent with the principles of Information Theory, data collected from

a single EEG channel significantly increases the complexity of sleep-state identification for humans and automated approaches. Therefore, the goal of this research has been to produce an open access and automated classifier that can reliably predict vigilance state based on a single cortical EEG from rodents. To that end, we have evaluated several of the commonly used Machine Learning techniques for suitability and success in this task. Our investigations include Decision Trees, Naive Bayes Classifiers, Random Forests, and Artificial Neural Networks. An Artificial Neural Network (ANN) has been developed to examine and ascertain the sleep state of each animal.

II. Methodology

A. Sleep EEG/EMG Data Collection

Adult Wistar rats (n=8) were used in experiments in a facility fully accredited by the American Association for the Accreditation of Laboratory Animal Care. Animals were kept on a 12/12 h light-dark cycle. All protocols were approved by the Institutional Animal Care and Use Committee at the University of South Carolina and were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals.

Rats were implanted with EEG/EMG telemetry devices (PhysioTel HD-S02, Data Science International, St. Paul, MN), as previously described [15]–[18]. Briefly, under isoflurane anesthesia, animals were placed in a stereotaxic frame. The transmitter device was intraperitoneally implanted through a dorsal incision of the abdominal region. After an incision at the midline of the head was made, EEG leads were secured to two surgical screws inserted into 0.5-mm burr holes at 2.0 mm anterior/1.5 mm lateral and 7.0 mm posterior/−1.5 mm lateral to bregma. Two EMG leads were inserted into the dorsal cervical neck muscle about 1.0 mm apart and sutured into place. The skin was sutured, and animals were recovered for a minimum of 7 days prior to experimentation. Sleep data were acquired in a quiet, designated room where rats remained undisturbed for the duration of recording using Ponemah 6.10 software (DSI). Digitized signal data were imported into NeuroScore 3.0 (DSI) and powerband frequencies (0 to 20 Hz) in 0.5 Hz increments were exported to CSV formatted flat files.

B. Data Processing and Annotation

Approximately 571 hours of continuously monitored electroencephalogram (EEG), electromyogram (EMG), and activity were recorded across the eight laboratory rodents without noise reduction. The collection of data was then partitioned into 10-second *epochs*, or segments, and labeled by a domain expert. Approximately 427 hours of the recorded 571 hours have been manually labeled by a domain expert. In this study, our focus has been based on the manually annotated 427 hours of the data. The remaining 144 hours of unlabelled data will be used in the future for more rigorous testing of our automated classification method. Based on the polysomnogram (PSG), an expert labels each 10-second epoch as one of three discrete classes: Paradoxical, Slow-wave, or Wake. Paradoxical sleep is also known as rapid-eye-movement (REM) sleep, because of the paradox of the high-frequency brain waves mimicking wakefulness despite being asleep. Slow-Wave sleep, characterized by low frequency, high amplitude brain waves, is also known as non-rapid

eye movement sleep (NREM), and it constitutes all sleep that is not REM sleep. Finally, the “wake” classification is given to a PSG that elicits characteristics of wakefulness. These sleep stages constitute all three *classes* and will be referred to as P, S, and W, which correspond to Paradoxical, Slow-wave, and Wake states, respectively.

42 discrete features were extracted from each 10-second epoch of continuous EEG, EMG, and activity signals. To obtain the first 40 features, the 10-second epoch is transformed from the time domain into the frequency domain using Discrete Fourier Transformation, then partitioned into 40 channels of equal width from 0 to 20 Hz. Thus, the first channel is the EEG from 0 to 0.5 Hz, the second channel is the EEG from 0.5 to 1 Hz, and so on. The 41st feature is that of the EMG, which is averaged over the 10-second epoch. The 42nd and final feature is the “Activity” feature, which is a derived parameter from Ponemah (indicating the level of an animal’s activity) which depends on the transmitter model, the speed with which the transmitter moves, outside radio interference, and variations from sensor to sensor.

C. Input Formalization

Many tactics in modern machine learning and artificial intelligence have been presented that aim to formalize the use of temporal data in the tasks of prediction and classification. Although the proper formulation of input can have a substantial impact on the trainability and the outcome of ML developments, in this investigation, we have implemented the most prevalent and natural approach. More specifically, the input to our ML activities consisted of a concatenation of five consecutive processed data (42 channels) that summarize a 10-second epoch from the raw continuous data. We postulate that a single epoch (summary of 10 seconds of data) is not the optimal temporal representation of time-series data for use in ML applications. Therefore, we choose to reformat the data such that each sample spans a longer period, theoretically providing each classifier with more temporal information and confidence. Five consecutive epochs encapsulate 50 seconds of the temporal signal, to which we will refer as the network input in the remainder of this report. Before this windowing, the data consisted of 154043 rows and 43 columns. Each row, which itself constitutes a 10-second epoch, consists of the 42 features and 1 output label describing the three stages of vigilance.

This mechanism of input data creation results in the following class distribution:

Total : 154039

P : 10028 (6.50% of total)

S : 64539 (41.77% of total)

W : 79472 (51.73% of total).

This data set was created by a simple moving window, where the first windowed sample will consist of samples 0–4 in the original data. The second windowed sample consisted of samples 1–5 in the original data, and the final windowed sample consisted of samples (n-5)-(n-1) in the original data. Since each input spans five individual vigilance states, various methods of arriving at a single output (given from five outputs) can be envisioned. In this work, we choose to label each windowed sample with the most frequent class in the

window. If there is a tie, we label the sample with the label of the 10-second epoch in the center. The total number of samples after the simple moving windowing has 4 samples less than the original data, which is well known to be the case when windowing data. Thus, each row in the input data consists of 210 features (5 sets of 42 features), and 1 label.

D. Data Balancing

The class distribution in the raw windowed data is unequal. Ideally, there would be an equal representation of each class, where each class constitutes 33% of the data, in this instance. However, class P is severely underrepresented at 6.51% of the total data, or 10028 samples. Classes S and W are over-represented at 41.90% and 51.59%, or 64539 samples and 79472 samples, respectively. In balancing the representation between classes, we aim to replicate copies of classes to the data while minimizing total samples. Minimizing total samples is important to improve training speed. By concatenating complete copies of any given class to itself, we aim to ensure that the model adequately generalizes to the entire class sample. Therefore, we concatenate 7 additional copies of the entire **P** class to the data, which produced the following and improved class distribution:

Total : 224235

P : 80224 (35.78% of total)

S : 64539 (28.78% of total)

W: 79472 (35.44% of total).

E. Training, Validation, and Testing Split and Shuffle

To finalize data preparation before training a classifier, the data must be shuffled and partitioned into training, validation, and testing sets. To perform the shuffling and partitioning, we use a function from the popular machine learning module in python scikit-learn. This function randomly shuffles the data and partitions it into training and testing data. We split the data into 80% training and 20% testing data. The only machine learning classifier which uses a validation set is the Artificial Neural Network. For the Artificial Neural Network, we further split the training data into 80% training, 20% validation, which constitutes 64% and 16% of the entire data. This process was performed only once to keep the training, testing, and validation (when needed) sets constant across each ML technique. A consistent training/testing set will help to establish a more consistent comparison of performances across multiple techniques while also reducing the data preparation time.

F. Evaluation of Machine Learning Classifiers

We propose to evaluate various machine learning techniques to classify vigilance states. Following are brief descriptions of Decision Trees, Random Forests, Naive Bayes Classifiers, Logistic Regression Classifiers, and Artificial Neural Networks. We aim to find the simplest model that achieves the highest accuracy.

G. Decision Tree

A Decision Tree Classifier is a supervised machine learning algorithm that performs classification tasks. The algorithm behind a Decision Tree makes a sequence of decisions

based on input features, one decision at each node along a path from the root to a leaf of the tree. The leaf node that the algorithm ends on for any given input determines the output class. Decision trees are self-interpretable, meaning the tree itself describes the underlying rules for classification. The advantages of the Decision Tree Classifier include its simplicity, interpretability, ability to model nonlinear data, ability to model high dimensional data, ability to work with large datasets to produce accurate results, and ability to handle outliers during training.

H. Random Forest

A Random Forest Classifier is a supervised predictive machine learning algorithm commonly used for classification tasks. A Random Forest consists of an ensemble of decision trees, each of which provides a "vote", or a classification, predicting class based on a majority of votes from the decision trees. Random forests generally outperform decision trees in terms of accuracy; however, the random forest is a *blackbox*, a model unable to describe its underlying rules for classification, sacrificing the interpretability of the decision tree.

I. Naive Bayes

A Naive Bayes Classifier is a supervised probabilistic machine learning algorithm commonly used for classification tasks. It relies on Bayes' Theorem, which is a theorem of conditional probabilities. Naive Bayes assumes strong independence between the input features. We use the Gaussian Naive Bayes Classifier based on the assumption that each input feature is normally distributed. The Naive Bayes Classifier is simple (and, therefore, computationally fast), scalable (requiring parameters linear in the number of features), and works well with high-dimensional data.

J. Logistic Regression

A Logistic Regression Classifier is a supervised predictive machine learning algorithm used for classification tasks. It is a type of generalized Linear Regression algorithm with a complex cost function. The cost function used here is the Sigmoid function, which continuously maps real-valued numbers between 0 and 1. We partition these continuously-mapped values from the cost function by various threshold values to determine output class. We include Logistic Regression in this work based on its speed, non-assumption of feature independence, and ubiquity in multi-class classification.

K. Artificial Neural Network

An artificial neural network is a supervised predictive machine learning technique that is successful in classification tasks in various applications [19]–[21]. A feed-forward neural network is a subset of neural networks in general where there are no cycles formed between neurons. We choose to use a specific type of feed-forward neural network, the multi-layer perceptron (MLP). An MLP consists of at least one hidden layer of neurons, as opposed to a single-layer perceptron, which does not have a hidden layer of neurons separate from the input and output layers. We propose to use a shallow neural network, as opposed to a deep

neural network, which is a subset of the multi-layer perceptron. For a neural network to be *shallow*, it means that the network has exactly one hidden layer with any number of neurons.

1) Architecture: For this investigation, we propose a fully connected artificial neural network with one hidden layer. The network has a single input layer where the number of neurons equals the number of input features and a single output layer where the number of neurons equals the number of classes. Thus, we use a neural network architecture similar to Figure.1 with 210 input neurons, 256 hidden layer neurons, and three output neurons. Non-linear activation functions allow neural networks to learn complex patterns. We use the rectified linear unit activation function after the hidden layer to introduce this non-linearity into the model. We use the softmax activation function after the output layer. Softmax is commonly used in multi-class classification problems, more general than sigmoid, and maps output into the range between 0 and 1, making it a good function for determining class. We use the *backpropagation* learning technique and the *adam* optimizer [22].

L. Evaluation of Model Performance

We evaluate the performance of each of the classifiers by comparing the predictions of each model to the manual scoring of domain experts. Testing accuracy is a ratio of the number of correctly classified samples by the model to the total number of samples. The primary measure of performance used in this investigation will be testing accuracy.

However, testing accuracy does not always fully describe the performance of a machine learning model. Oftentimes it is important to maximize the true positive classifications or to minimize the false positive classifications. In order to measure these situations, we have the proportions called *precision* and *recall*. *Precision* is the proportion of model classifications that correspond to the sample's true class. *Recall* is the proportion of samples that are actually classified by the model correctly. In addition to these metrics, we also provide *F1 Score*. The *F1 Score* is the harmonic mean of precision and recall. Finally, *AUC*, or Area Under the receiver operating characteristic Curve, represents the probability that the model ranks a random positive example higher than a random negative example. We choose to display testing classifications for each machine learning model in the form of a Confusion Matrix. In each confusion matrix, class 0 corresponds to class P, class 1 corresponds to class S, and class 2 corresponds to class W. Accuracy, precision, recall, F1 score, and AUC can all be derived from a confusion matrix.

III. Results

We evaluate the performance of the aforementioned methodologies primarily through testing accuracy. The testing accuracies for each methodology are described in Table. I. F1 Score and AUC Score are provided for each machine learning technique; however, it will suffice to focus solely on accuracy.

Notably, the highest performing classifier is the Random Forest model with a testing accuracy of 95.78%. The confusion matrix for the testing classification of this model is shown in Fig. 2. Other metrics such as precision and recall may be derived from the confusion matrix. The model performs excellently on class 0, which corresponds to

paradoxical sleep. The most frequently misclassified prediction-label pair occurred between slow-wave sleep and wakefulness. There were 1017 instances of predicting wakefulness where the true label was slow-wave and 606 samples where the model predicted slow-wave sleep where the true label was wake.

The second-highest performing classifier is the Artificial Neural Network with a testing accuracy of 93.31%. Across the entire testing partition, the model achieves over 93% categorical accuracy, which is comparable or higher than many other methodologies in the literature [8], [23], [24]. The confusion matrix for the testing classification of this model is shown in Fig. 6. The prediction-label pair that was most frequently misclassified by the ANN occurred when the model predicted slow-wave and the true label was wake. Interestingly, this is the same confusion pair that occurred most frequently in the Random Forest model. Theoretically, there is a dimension that would reliably distinguish between slow-wave and wake. Future research will investigate this relationship and attempt to distinguish more reliably and accurately between these two classes. The ANN achieved a high F1 Score and AUC Score of 93.9% and .9867, respectively. After training for only 50 epochs, the model achieves a testing categorical accuracy of 94.01%, testing precision of 94.54%, testing recall of 93.43%, and a testing AUC of 99%. The value of these metrics over the period of training are shown in Fig. 7, Fig. 8, Fig. 9, and Fig. 10. Not only did this model achieve a high degree of categorical accuracy, precision, and recall, but it did this with very little training and computation. All training instances consisted of only 50 epochs, each of which took approximately 220.66 seconds when run on a 2.4 GHz 7th Generation Intel(R) Core(T M) i7–7700HQ Quad-Core Processor with 16GB (2×8GB) DDR4 at 2400Mhz.

The Decision Tree model performed well with an accuracy of 92.77%; however, we do not discuss this model any further because it is a special case of the Random Forest. Logistic Regression and Naive Bayes had testing accuracies of 77.33% and 74.37%, respectively. These low accuracies suggest that this vigilance state classification problem is not trivial and that the accuracies reached by the Random Forest and the Artifical Neural Network are remarkable achievements.

IV. Conclusion

Many works which rely on multiple EEG channels have successfully produced automated vigilance state classifiers; however, the hardware required to obtain such signals produces undesirable disadvantages. Here we evaluate various machine learning techniques in vigilance state classification based on a single EEG channel. Random Forests and Artificial Neural Networks produced remarkable accuracies of approximately 96% and 93%. In evaluation of these classification techniques against human scoring as ground truth, we note that humans may make misclassifications. In fact, due to the rigidly patterned nature of the data and the power of these statistical models, it would be appropriate to reevaluate human scores based on model classifications. Future research will have a domain expert label polysomnograms from the additional 147 hours of unscored data aided by the model from this work. Additionally, future research will have a domain expert reevaluate classification pairs that confused the models in this work. Here we achieved two accurate and reliable

models that can be used immediately in an automated vigilance state classifier and may be reinforced in future work.

References

[1]. Cirelli C and Tononi G, "Is sleep essential?" Journal Name Here, vol. 6, no. 8, p. e216, Aug. 2008. [Online]. Available: 10.1371/journal.pbio.0060216

[2]. Watson NF, Badr MS, Belenky G, Bliwise DL, Buxton OM, Buysse D, Dinges DF, Gangwisch J, Grandner MA, Kushida C, Malhotra RK, Martin JL, Patel SR, Quan S, and Tasali E, "Recommended amount of sleep for a healthy adult: A joint consensus statement of the american academy of sleep medicine and sleep research society," Journal Name Here, Jun. 2015. [Online]. Available: 10.5665/sleep.4716

[3]. and Watson Nathaniel F., Badr MS, Belenky G, Bliwise DL, Buxton OM, Buysse D, Dinges DF, Gangwisch J, Grandner MA, Kushida C, Malhotra RK, Martin JL, Patel SR, Quan SF, Tasali E, Twery M, Croft JB, Maher E, Barrett JA, Thomas SM, and Heald JL, "Joint consensus statement of the american academy of sleep medicine and sleep research society on the recommended amount of sleep for a healthy adult: Methodology and discussion," Journal Name Here, vol. 38, no. 8, pp. 1161–1183, Aug. 2015. [Online]. Available: 10.5665/sleep.4886

[4]. Elmer GI, Brown PL, and Shepard PD, "Engaging research domain criteria (RDoC): Neurocircuitry in search of meaning," Schizophrenia Bulletin, vol. 42, no. 5, pp. 1090–1095, Jul. 2016. [Online]. Available: 10.1093/schbul/sbw096 [PubMed: 27412648]

[5]. Missig G, Mokler EL, Robbins JO, Alexander AJ, McDougle CJ, and Carlezon WA, "Perinatal immune activation produces persistent sleep alterations and epileptiform activity in male mice," Neuropsychopharmacology, vol. 43, no. 3, pp. 482–491, Oct. 2017. [Online]. Available: 10.1038/npp.2017.243 [PubMed: 28984294]

[6]. Gao V, Turek F, and Vitaterna M, "Multiple classifier systems for automatic sleep scoring in mice," J Neurosci Methods, vol. 264, pp. 33–39, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26928255 [PubMed: 26928255]

[7]. Gross BA, Walsh CM, Turakhia AA, Booth V, Mashour GA, and Poe GR, "Open-source logic-based automated sleep scoring software using electrophysiological recordings in rats," J Neurosci Methods, vol. 184, no. 1, pp. 10–8, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19615408 [PubMed: 19615408]

[8]. Miladinovic D, Muheim C, Bauer S, Spinnler A, Noain D, Bandarabadi M, Gallusser B, Krummenacher G, Baumann C, Adamantidis A, Brown SA, and Buhmann JM, "Spindle: End-to-end learning from eeg/emg to extrapolate animal sleep scoring across experimental settings, labs and species," PLoS Comput Biol, vol. 15, no. 4, p. e1006968, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30998681 [PubMed: 30998681]

[9]. Stephenson R, Caron AM, Cassel DB, and Kostela JC, "Automated analysis of sleep-wake state in rats," J Neurosci Methods, vol. 184, no. 2, pp. 263–74, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19703489 [PubMed: 19703489]

[10]. Tredger JM, Grevel J, Naoumov N, Steward CM, Niven AA, Whiting B, and Williams R, "Cyclosporine pharmacokinetics in liver transplant recipients: evaluation of results using both polyclonal radioimmunoassay and liquid chromatographic analysis," Eur J Clin Pharmacol, vol. 40, no. 5, pp. 513–9, 1991. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/1884727 [PubMed: 1884727]

[11]. Louis RP, Lee J, and Stephenson R, "Design and validation of a computer-based sleep-scoring algorithm," J Neurosci Methods, vol. 133, no. 1–2, pp. 71–80, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14757347 [PubMed: 14757347]

[12]. Ellen JG and Dash MB, "An artificial neural network for automated behavioral state classification in rats," PeerJ, vol. 9, p. e12127, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/34589305 [PubMed: 34589305]

[13]. Kramer K and Kinter LB, "Evaluation and applications of radiotelemetry in small laboratory animals," Physiol Genomics, vol. 13, no. 3, pp. 197–205, 2003. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12746464 [PubMed: 12746464]

[14]. Papazoglou A, Lundt A, Wormuth C, Ehninger D, Henseler C, Soos J, Broich K, and Weiergraber M, "Non-restraining eeg radiotelemetry: Epidural and deep intracerebral stereotaxic eeg electrode placement," J Vis Exp, no. 112, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27404845

[15]. Rentschler KM, Baratta AM, Ditty AL, Wagner NTJ, Wright CJ, Milosavljevic S, Mong JA, and Pocivavsek A, "Prenatal kynurenine elevation elicits sex-dependent changes in sleep and arousal during adulthood: Implications for psychotic disorders," Schizophr Bull, vol. 47, no. 5, pp. 1320–1330, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33823027 [PubMed: 33823027]

[16]. Baratta AM, Buck SA, Buchla AD, Fabian CB, Chen S, Mong JA, and Pocivavsek A, "Sex differences in hippocampal memory and kynurenic acid formation following acute sleep deprivation in rats," Sci Rep, vol. 8, no. 1, p. 6963, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29725029 [PubMed: 29725029]

[17]. Pocivavsek A, Baratta AM, Mong JA, and Viechweg SS, "Acute kynurenine challenge disrupts sleep-wake architecture and impairs contextual memory in adult rats," Sleep, vol. 40, no. 11, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29029302

[18]. Baratta AM, Kanyuch NR, Cole CA, Valafar H, Deslauriers J, and Pocivavsek A, "Acute sleep deprivation during pregnancy in rats: Rapid elevation of placental and fetal inflammation and kynurenic acid," Neurobiol Stress, vol. 12, p. 100204, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32258253 [PubMed: 32258253]

[19]. Cole CA, Anshari D, Lambert V, Thrasher JF, and Valafar H, "Detecting smoking events using accelerometer data collected via smartwatch technology: Validation study," JMIR Mhealth Uhealth, vol. 5, no. 12, p. e189, Dec. 2017. [Online]. Available: 10.2196/mhealth.9035 [PubMed: 29237580]

[20]. Fawcett TM, Irausquin SJ, Simin M, and Valafar H, "An artificial neural network approach to improving the correlation between protein energetics and the backbone structure," Proteomics, vol. 13, no. 2, pp. 230–238, Dec. 2012. [Online]. Available: 10.1002/pmic.201200330 [PubMed: 23184572]

[21]. Odhiambo CO, Cole CA, Torkjazi A, and Valafar H, "State transition modeling of the smoking behavior using lstm recurrent neural networks," 2020.

[22]. Kingma DP and Ba J, "Adam: A method for stochastic optimization," 2017.

[23]. Exarchos I, Rogers AA, Aiani LM, Gross RE, Clifford GD, Pedersen NP, and Willie JT, "Supervised and unsupervised machine learning for automated scoring of sleep–wake and cataplexy in a mouse model of narcolepsy," Sleep, vol. 43, no. 5, Nov. 2019. [Online]. Available: 10.1093/sleep/zsz272

[24]. Yamabe M, Horie K, Shiokawa H, Funato H, Yanagisawa M, and Kitagawa H, "MC-SleepNet: Large-scale sleep stage scoring in mice by deep neural networks," Scientific Reports, vol. 9, no. 1, Oct. 2019. [Online]. Available: 10.1038/s41598-019-51269-8

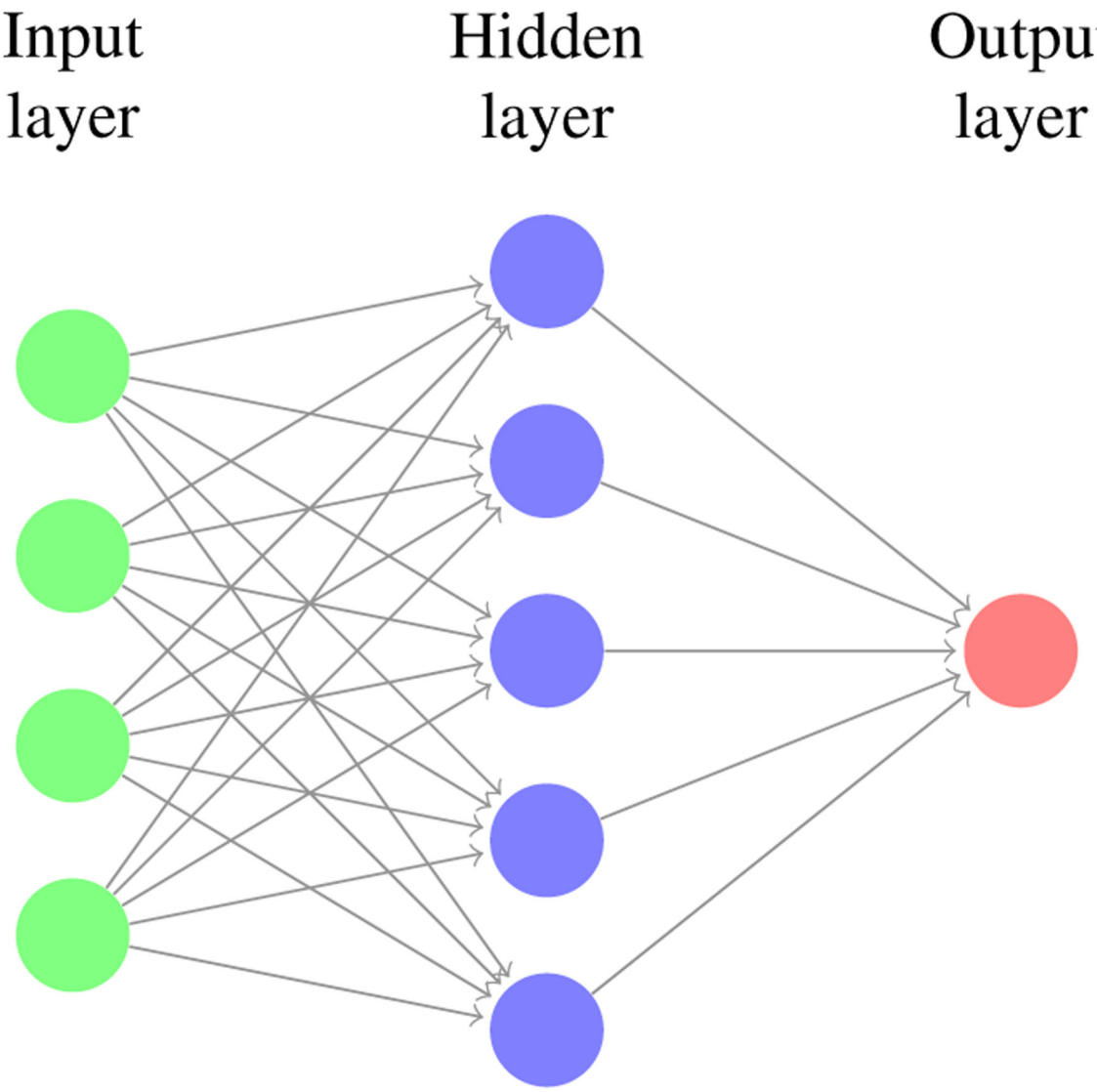

Fig. 1.

Fully-connected artificial neural network with shape (210,256,3), meaning the input layer has 210 neurons, the single hidden layer has 256 neurons, and the output layer has 3 neurons.

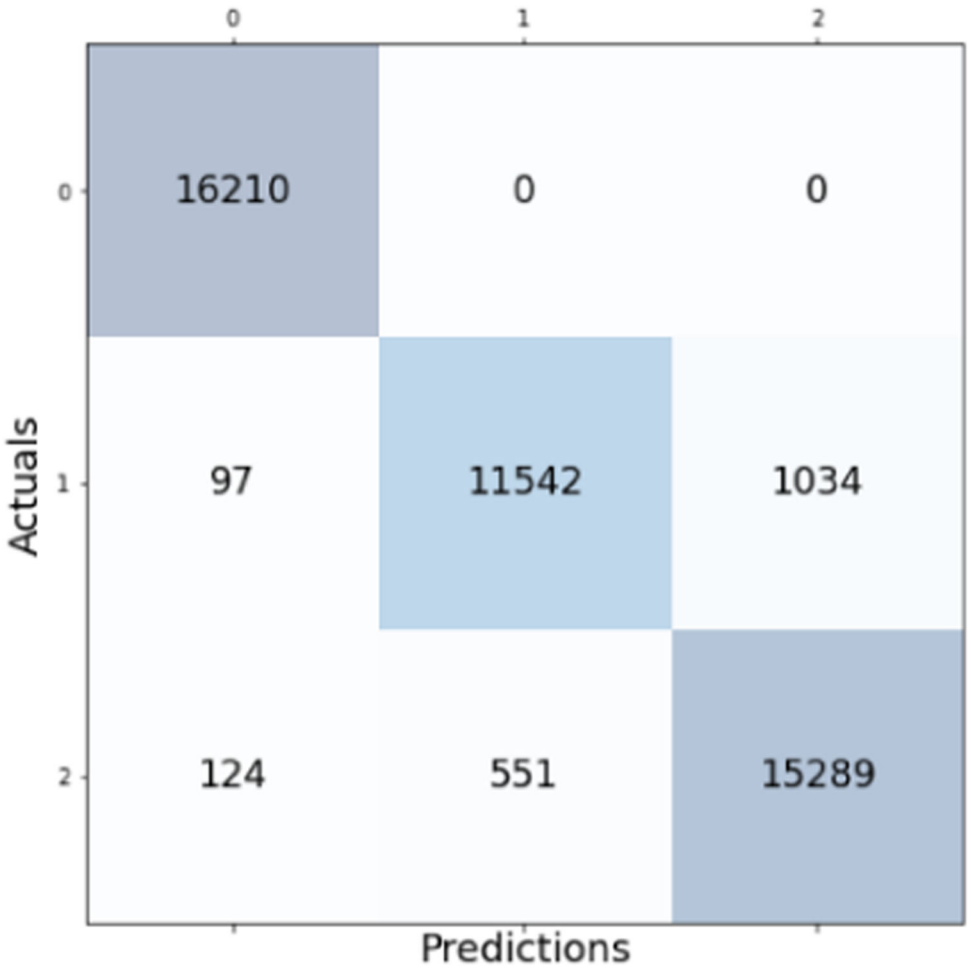

Fig. 2.
Confusion matrix for the Random Forest model with overall accuracy of 95.78%.

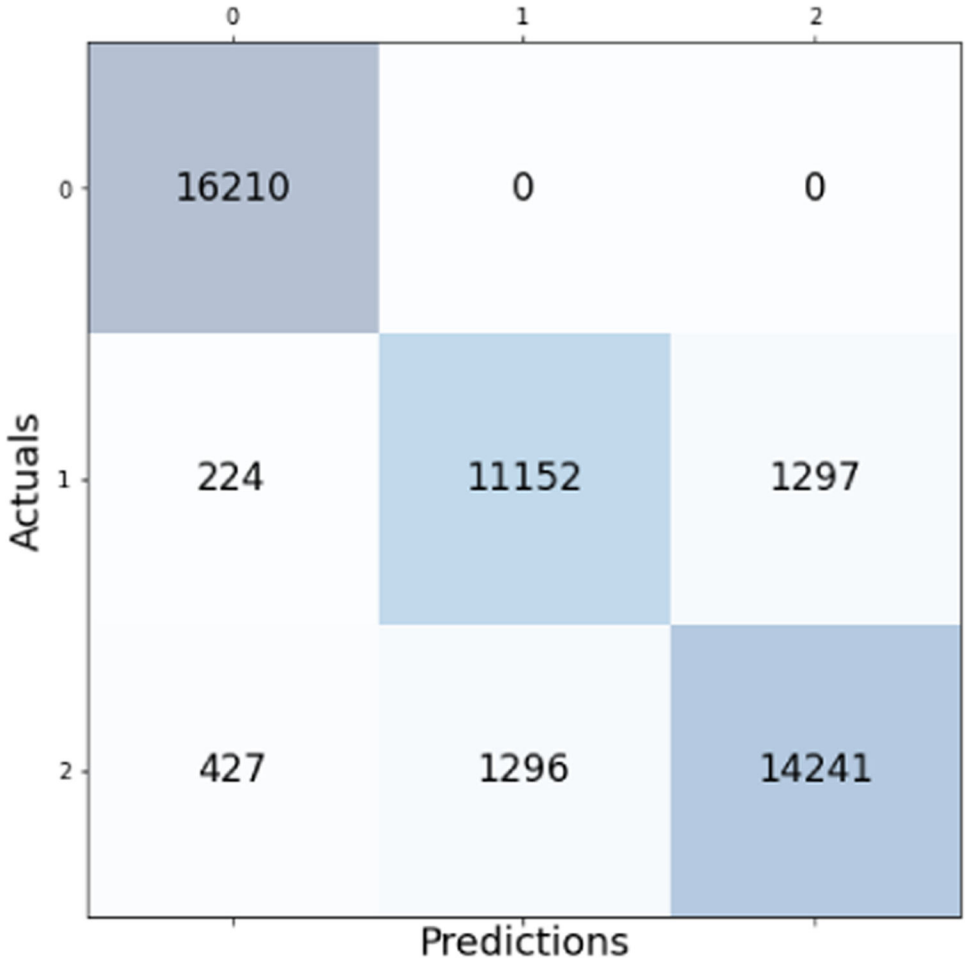

Fig. 3.
Confusion matrix for the Decision Tree model with overall accuracy of 92.77%.

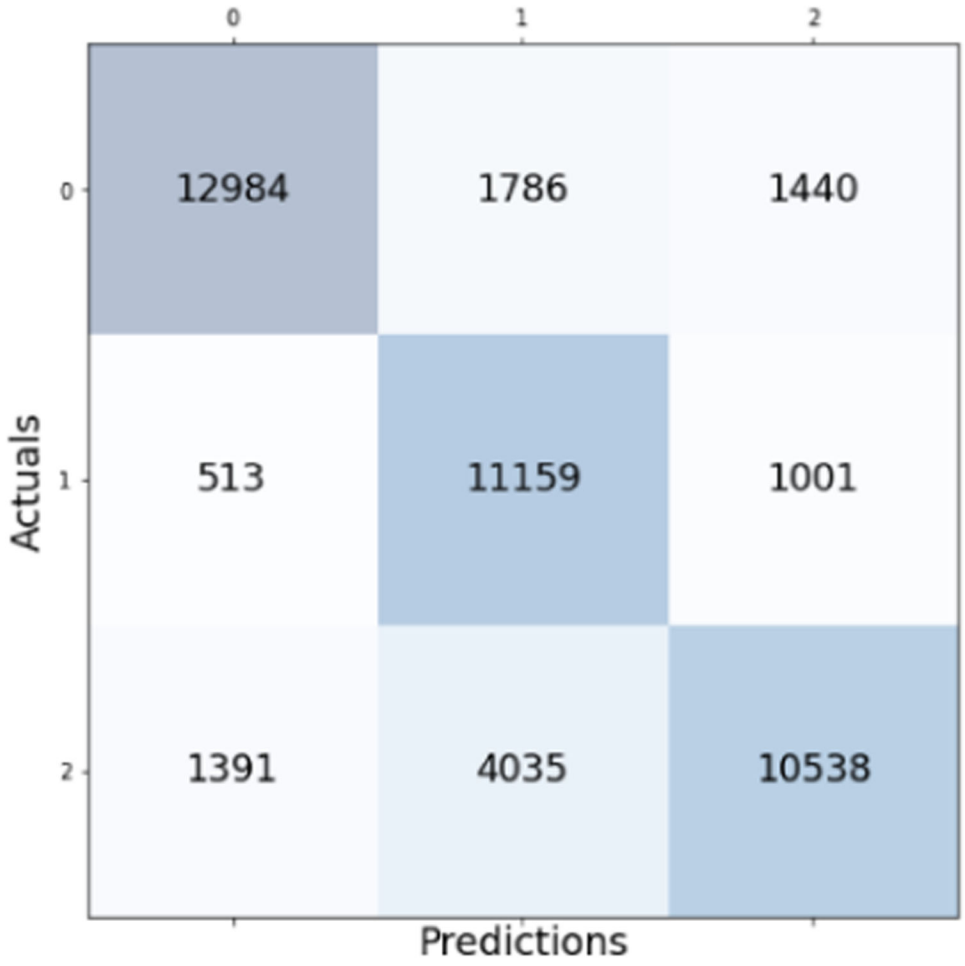

Fig. 4.
Confusion matrix for the Logistic Regression model with overall accuracy of 77.33%.

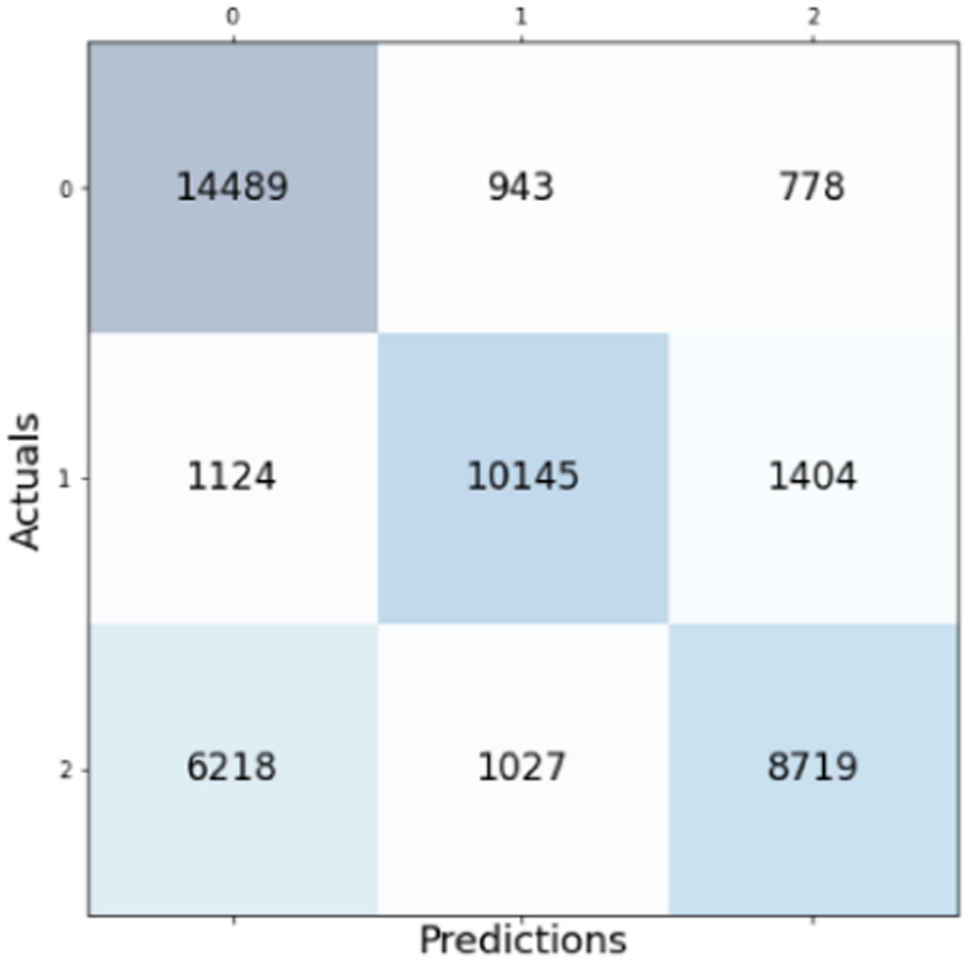

Fig. 5.
Confusion matrix for the Naive Bayes model with overall accuracy of 74.37%.

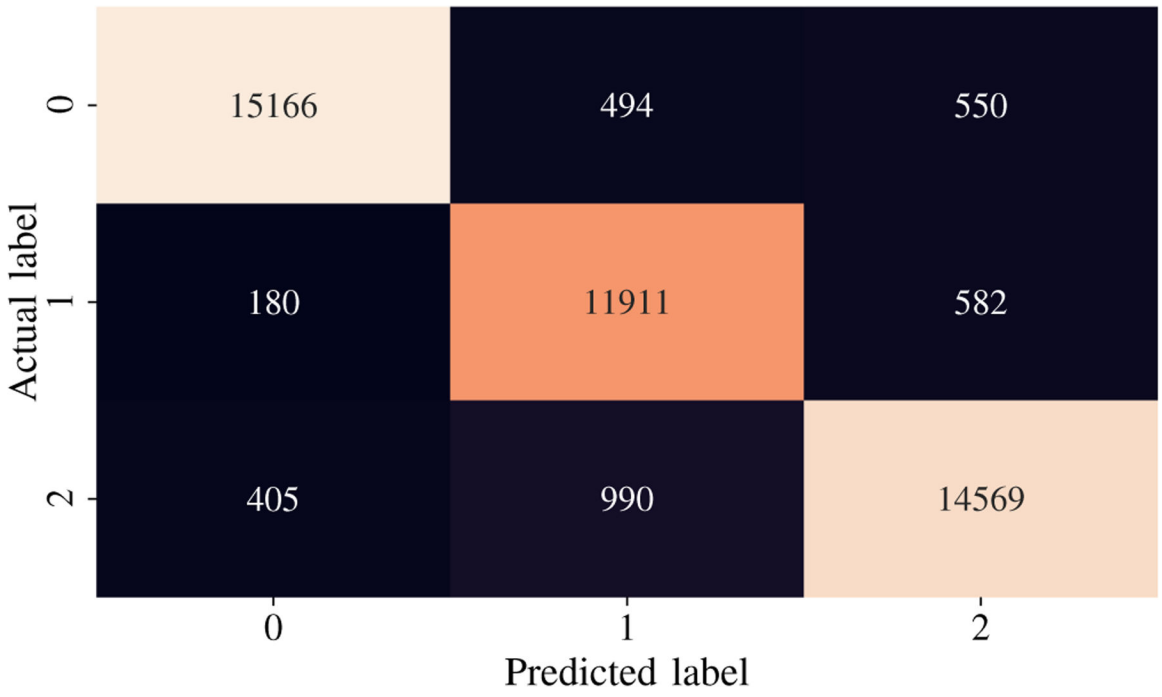

Fig. 6.
Confusion matrix for the Artificial Neural Network classifier with overall accuracy of 93.31%.

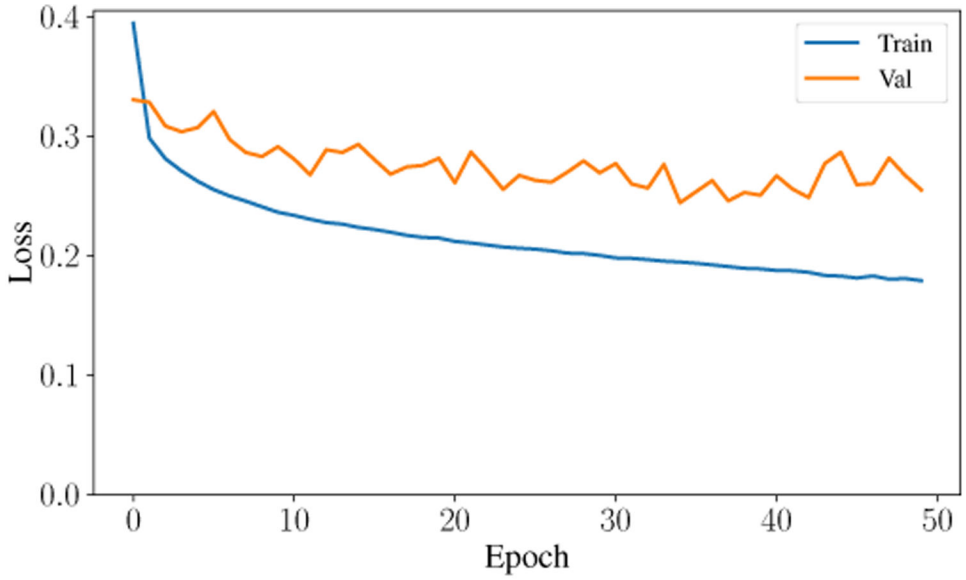

Fig. 7.
Loss curve for training and validation data over 50 epochs.

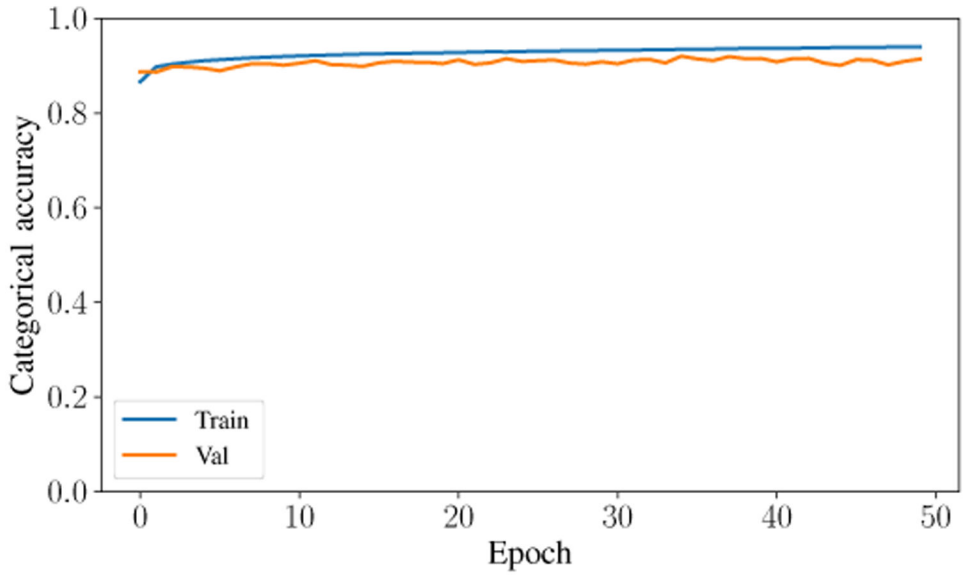

Fig. 8.
Accuracy curve for training and validation data over 50 epochs.

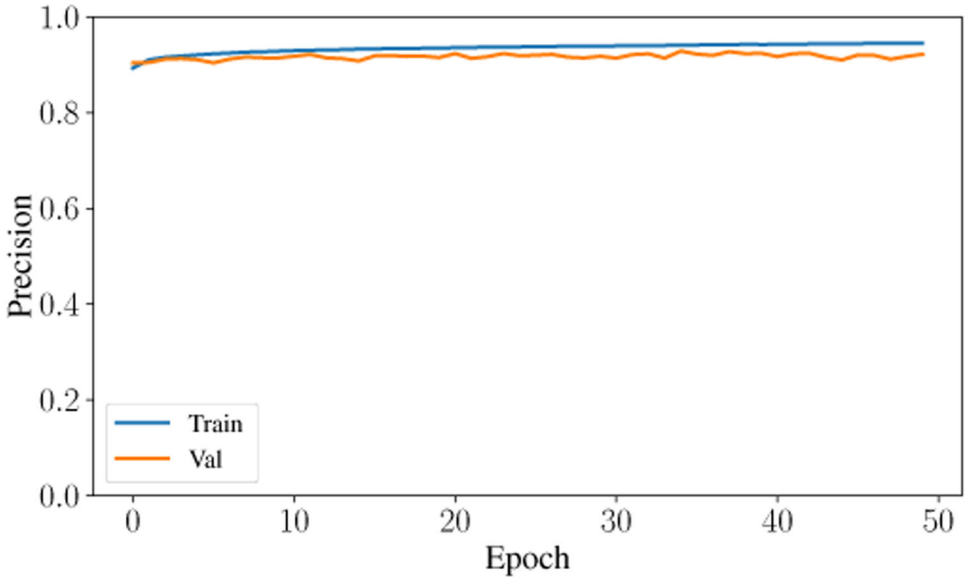

Fig. 9.
Precision curve for training and validation data over 50 epochs.

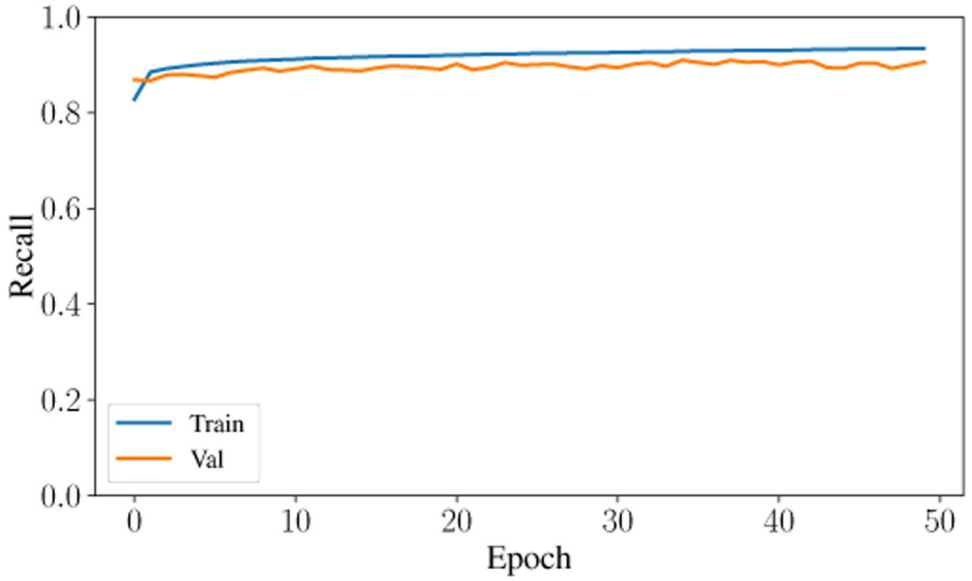

Fig. 10.
Recall curve for training and validation data over 50 epochs.

Table I

Performance of various machine learning models in sleep stage classification.

Classifier	Accuracy	FI Score	AUC Score
Random Forest	95.78%	96%	.9954
ANN	93.31%	93.9%	.9867
Decision Tree Classifier	92.77%	93%	.9427
Logistic Regression	77.33%	77%	.9242
Naive Bayes	74.37%	74%	.9011