Scientific Journal of Informatics
Vol. 11, No. 4, Nov 2024
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Implementation of K-Nearest Neighbor in Case-Based Reasoning for Mental Health Diagnosis Systems

Ardian Pamungkas[1*], R Rizal Isnanto[2], Dinar Mutiara Kusumo Nugraheni [3],
[1]Department of Information System, Universitas Diponegoro, Indonesia
[2]Department of Computer Engineering, Universitas Diponegoro, Indonesia
[3]Department of Informatics, Universitas Diponegoro, Indonesia

Abstract.
Purpose: Assessing a model that employs the K-Nearest Neighbor (KNN) technique within Case-Based Reasoning (CBR) for diagnosing mental health disorders, concentrating on conditions such as anxiety, depression, stress, and normalcy, while enhancing its efficacy through the utilization of historical case data for more accurate and tailored diagnostic suggestions.
Methods: This study implements the KNN method in CBR to create a mental health diagnosis system that can provide accurate results without the need for complex models or intensive training. This method effectively addresses various patient needs by utilizing previous case data to provide a personalized and case-based diagnosis. This system is designed to tackle mental health issues like anxiety, depression, and academic stress, utilizing a case study of students from ITBK Bukit Pengharapan.
Result: This study developed a KNN-based model for mental health diagnosis, achieving 84.62% accuracy on test data. Data processing techniques like text mining, oversampling, and cosine similarity improved performance. With an optimal K value of 2, the model achieved 88% precision, 85% recall, and an F1-score of 84%. The anxiety label performed perfectly, with 100% precision, recall, and F1-score.
Novelty: This study adds innovation by integrating the rarely used CBR and KNN algorithms for mental health diagnosis systems. Innovative techniques like text mining, oversampling to get around data integration, and cosine similarity computations, which greatly enhance model performance, assist this strategy. Because this method improves accuracy and expedites the diagnosis process, both of which support clinical decision-making, it may be able to help mental health professionals.

Keywords: CBR, Diagnosis, K-nearest neighbor, Mental health
Received January 2024 / **Revised** February 2025 / **Accepted** February 2025

INTRODUCTION

Mental health diagnostic systems are gaining significance in the rapidly changing digital age, offering psychological support and counseling to individuals in need, regardless of their geographic location. [1]. The application of technology in the field of mental health, such as diagnosis systems, can bring benefits by increasing wider service accessibility and providing more space to help facilitate health diagnosis [2]. The mental health diagnosis system is an innovative solution to provide wider accessibility to mental health diagnosis services [3]. This study utilized the K-Nearest Neighbor (KNN) method within case-based reasoning (CBR) to develop a mental health diagnosis system, with the goal of enabling patient diagnosis and counseling without geographical or temporal limitations. The KNN method was chosen for this research due to its simplicity, as it does not necessitate a complex model and can readily adapt to new data without requiring extensive training. [4]. If the training data is representative, KNN can provide accurate results [5]. Compared to other methods that require complex mathematical calculations [6], larger computing resources, and longer training times, KNN adapts quickly to new data without requiring expensive retraining [7].

At present, counselors face a diverse range of patient needs, with many individuals lacking the essential knowledge and skills in mental health education. This deficiency contributes to a generally low level of mental health education overall. [8]. Considering the limited availability of counselors and the wide variety of patients receiving treatment, remote mental health diagnosis can help address each patient's specific

*Corresponding author.
Email addresses: ardian.pamungkas05@gmail.com (Pamungkas)
DOI: 10.15294/sji.v11i4.19912

issues. By incorporating the KNN method within case-based reasoning (CBR) in a mental health diagnosis system, it offers a more personalized and accurate approach to mental health diagnoses based on existing cases. [9] [10]. The urgency in this study lies in increasing the effectiveness of mental health diagnosis, ensuring that each individual receives support that is appropriate to their conditions and needs. This study anticipates that this approach will enhance the efficiency and effectiveness of mental health diagnosis. The goal of this study is to evaluate the viability of a method for use in a mental health diagnosis system. This study aims to integrate the KNN method into the CBR system to improve the precision and accuracy of diagnostic recommendations by combining the two approaches. The choice of the KNN method within the CBR framework was driven by its capability to process and analyze historical case data, enabling it to deliver suitable diagnoses. CBR will help in finding a diagnosis based on similarities with previous cases [11]. A database, lacking rules but containing a series of encountered and resolved problems or events, serves as the fundamental source of knowledge for the CBR method. This research solves new problems by identifying the most similar problems and their potential adaptations [12].

Previous research on CBR and KNN has been used to perform classification in health, particularly in mental health diagnosis [13], [14]. CBR makes decisions based on previous cases as a reference to obtain a diagnosis for new cases. However, after obtaining a diagnosis, users are expected to consult further with a doctor to get a definitive diagnosis, as this expert system is designed only for initial consultation. Regarding KNN, it demonstrates that the early stress detection system can accurately identify stress. Furthermore, the system exhibits rapid computation time, highlighting one of the benefits of incorporating KNN in this context.

KNN will employ data from its closest neighbors to deliver more precise and prompt diagnostic suggestions [15]. The basic theory of KNN is that in a calibration dataset, KNN finds a group of k samples that are closest to the unknown sample [16]. For these k samples estimate the unknown sample by taking the mean of the response variables [17]. This research will explore the implementation of KNN within CBR as a decision support system for diagnosing mental health issues, specifically focusing on limitations related to conditions such as anxiety, depression, and academic stress. The case study will involve students from ITBK Bukit Pengharapan, and it will also address the constraints on the types of data and information utilized in the development of the expert knowledge base system.

This research aims to determine the model's performance for diagnosing mental health ability using CBR where the method used is KNN. Moreover, it aims to explore if implementing KNN in CBR may enhance the research results. This study's advantages are assessing the efficiency of the model in identifying the mental health symptoms using KNN in CBR and the second is minimizing how many resources are applied to enhance the workflow and customization of mental health detection recommendation services of previous case data and machine learning algorithms. Furthermore, this research offers empirical evidence regarding the effectiveness of KNN within CBR, thereby enriching academic literature and technology. It also supports the advancement of more responsive and adaptive mental health diagnostic systems.

METHODS

This study utilizes input data consisting of mental health cases, such as anxiety, depression, academic stress, normal, and other conditions. This input data is obtained from distributing questionnaires to students living in dormitories and FGDs with Counselors, which will then undergo the Pre-processing stage. The collected data is then validated by the counselor where the purpose of this validation is to ensure that the results of the questionnaire distribution are truly in accordance with the needs of the analysis in this study.

The results of the questionnaire validation are applied to the data preprocessing process using Text Mining Techniques, which include Tokenizing, Filtering, Stemming, TF-IDF, Cosine Similarity to ensure data consistency and quality. Subsequently, the data is divided into Training Data and Testing Data. In this study, the data was divided into two groups: 80 % for training data and 20 % for testing. Then, from 80% of the training data, it is further divided by 20% for validation data. Training Data is utilized to build and train the KNN model in CBR, while Testing Data is employed to evaluate the performance of the trained model. Figure 1 illustrates the sequence of stages outlined by the research methodology outlines for the study.

Validation of questionnaire by counselor
The objective of validating questionnaire data results with counselors is to ensure high-quality data for mental health diagnosis research. Subsequently, the questionnaire data that meets the criteria will be assessed by assigning labels used in the study, namely stress, anxiety, depression, and normal.

Data partitioning
Data partition stage divides the data into two main groups: training data (80%) and testing data (20%) for [18]. The subsequent step involves further partitioning of the training data into 20% for validation data. The purpose of validation data is to evaluate interim results as the model evolves by testing the model on a small portion of data not utilized for major training iteration. It is supposed to enhance the performance and generalization of the model in distinguishing among different types of patterns for unknown data. This segmentation is expected to increase the model evaluation and accuracy for handling complex real cases of mental health problems of students.

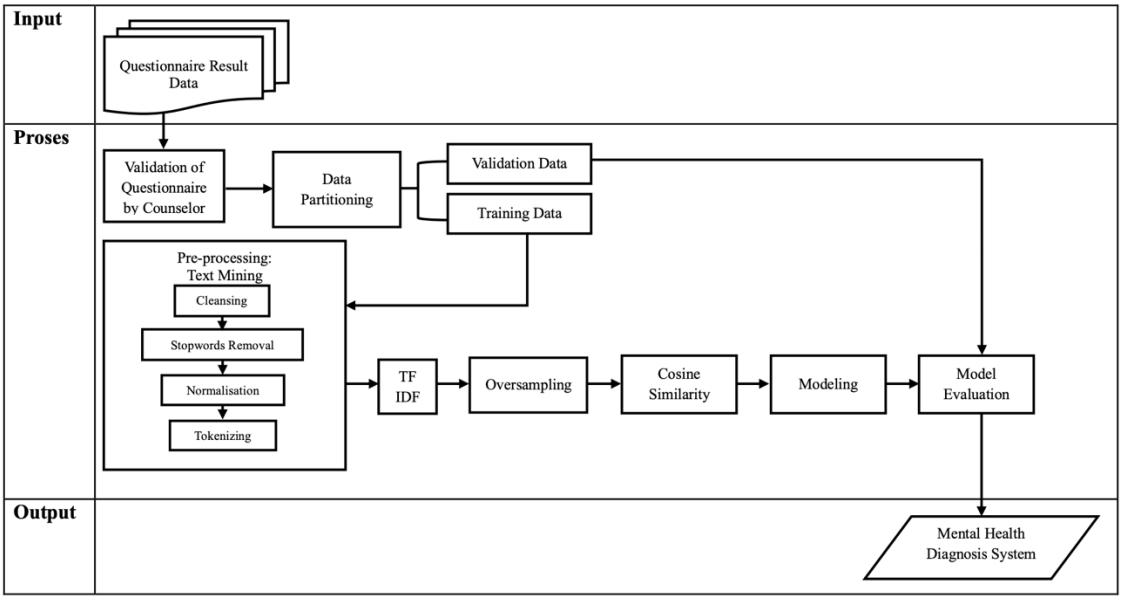

Figure 1. Information system framework

Pre-processing
This preprocessing is aimed at organizing raw data for further analysis [19]. This article outlines the preprocessing pipeline, including cleaning, stop-word removal, normalization, and tokenization.

Data cleaning
This process involves removing irrelevant letters or symbols from textual data, such as punctuation, numbers, and other unnecessary special characters that do not contribute to the research. Data cleaning is the procedure through which data is refined for application in machine learning models. [20].

Stopword removal
This step consists of removing the typical phrases which lack any substantial meaning in the analysis such as "and," "or," "is," etc. The stopword removal reduces noise in data and improves the efficiency of processing [21].

Normalization
Normalization is a part of the process of standardizing the text in which all the letters are converted into lowercase so that the differences of upper and lower cases will not cause any difference in the analysis [22].

Tokenization
Tokenization is the process of segmenting text into smaller units of words. This facilitates further lexical analysis and aids in feature extraction from the text [23].

Upon completion of these steps, the text data will be more organized and ready for use in machine learning models.

TF-IDF

The TF-IDF (Term Frequency-Inverse Document Frequency) technique in text processing assesses a word's importance in a document in relation to other words in the document set [24]. TFG-IDF consists of two metrics: Word Frequency (TF): measures how often a word appears in a document; the more frequently a word appears, the higher the TF value. Inverse Document Frequency (IDF): measures how important a word is. A low IDF value for words that appear frequently in many documents is the result of multiplying the logarithm of the total number of documents divided by the number of documents containing the word. By multiplying TF by IDF, TF-IDF gives higher weight to words that appear frequently in one document but rarely in others, thus helping to identify more relevant words for analysis. Information retrieval and text mining often use this technique for feature engineering and text classification.

Oversampling

To tackle the issue of data imbalance, the Synthetic Minority Oversampling Technique (SMOTE) creates new synthetic data for the minority class by utilizing the positions of existing minority data points.. SMOTE reduces the overfitting problem that often occurs with random oversampling techniques [25]. Several studies have used the SMOTE method to address data imbalance.

Cosine similarity

The cosine similarity method calculates the cosine of the angle between two vectors in a vector space and produces a similarity value that ranges from zero to one [17]. Values approaching one signify a greater degree of similarity between documents, whereas values further from zero indicate a lower degree of similarity. This method is highly effective as it normalizes the length of the vectors, enabling precise measurements of similarity between documents based on word frequency.

Modeling

The next step involves applying the Case-Based Reasoning (CBR) [26] model and implementing the KNN algorithm [27] for case retrieval, which matches new cases with existing cases in the database based on vector similarity. During the reuse stage, the model will adopt the diagnosis from the most similar case and make necessary adjustments. In the revise stage, additional testing and modifications of the diagnosis will occur. Finally, in the retain stage, the system will save the adjusted diagnosis back into the database. This process enables the system to deliver accurate, case-based recommendations for addressing students' mental health issues.

Model evaluation

According to the statement put forward, it can be appreciated that multiclass confusion matrix aims to provide the user with an understanding of the applicability of the model after being used [28]. With this matrix, it can be seen how well the model is able to distinguish between positive examples (cases diagnosed with mental health problems) and negative examples (other cases). Accuracy, recall, precision, and the F1 value can all be derived from the multiclass confusion matrix results, providing a comprehensive overview of the model’s performance. This evaluation makes it easier to evaluate the model’s efficiency in making specialist diagnoses and identifying areas for improvement the model would need to address.

Development of a mental health diagnosis system

After the model evaluation, the system can classify new cases and provide all the important information for further diagnosis and counseling interventions. The design is user-friendly, aiming to present results clearly and understandably when used for model testing. With this visual interface, users can evaluate the model's effectiveness of the model and see the results of mental health condition classification.The approach ensures the system's accuracy in categorizing results and its practicality for testing and model evaluation. This application is developed using Flask or Streamlit for the web interface, Scikit-learn for the text classification prediction model, and Pandas and NumPy for data processing. This system is a Python-based system involves predictions using machine learning models and a web-based user interface.

RESULTS AND DISCUSSIONS

The distributed questionnaire data set consists of 238 rows of student mental health cases. A clinical psychologist labels each row of data with a mental health condition column. Out of the 238 mental health conditions, there are 35 related to anxiety, 48 associated with depression, 93 classified as normal, and 62 linked to stress. The data will then be divided into 80% for training and 20% for testing. Furthermore, the training data further divided by 20% as validation data. The results of the data partitioning are displayed in Table 1.

Table 1. Partitioning data

Label Data	Training Data	Validation Data	Testing Data
Anxiety	22	3	10
Depression	32	8	8
Normal	64	12	17
Stress	34	15	13
Total Data:	**152**	**38**	**48**

Text mining will perform pre-processing on the training data, specifically data cleaning, with the goal of eliminating irrelevant elements or disturbing analysis. Next, stopwords will be removed to eliminate irrelevant elements, and the data will be normalized by converting it to lowercase. Following this, tokenization will be performed to break sentences into individual words or phrases. This pre-processed text allows the model to operate with more consistent and representative data, ultimately improving accuracy.

Table 2. Data pre-processing result

Case Before Pra-Processing	Biasa kena masalah, hanya bisa pasrah dan berdoa. Pada saat UTS dan UAS, dikejar tagihan DPA, berdoa dan bekerja dengan mood yang tidak menentu.
After Cleaning	Biasa kena masalah hanya bisa pasrah dan berdoa pada saat uts dan uas dikejar tagihan dpa berdoa dan bekerja dengan mood yang tidak menentu
After Stopword Removal	Biasa kena masalah hanya pasrah berdoa saat uts uas kejar tagihan dpa berdoa bekerja mood tidak menentu
After Normalization	biasa kena masalah hanya pasrah berdoa saat uts uas kejar tagihan dpa berdoa bekerja mood tidak menentu
After Tokenizing	"['biasa', 'kena', 'masalah', 'hanya', 'pasrah', 'berdoa', 'saat', 'ut', 'ua', 'kejar', 'tagihan', 'dpa', 'berdoa', 'bekerja', 'mood', 'tidak', 'menentu']"

The next step involves assessing the importance of a word in a larger document using the TF-IDF process. Figure 2 shows the results of the TF-IDF calculation.

```
Data setelah perhitungan TF-IDF:
   abang  acara  ada  adanya  adaptasi  adik  administrasi      aduk  agak  \
0    0.0    0.0  0.0     0.0       0.0   0.0           0.0  0.071361   0.0
1    0.0    0.0  0.0     0.0       0.0   0.0           0.0  0.000000   0.0
2    0.0    0.0  0.0     0.0       0.0   0.0           0.0  0.000000   0.0
3    0.0    0.0  0.0     0.0       0.0   0.0           0.0  0.000000   0.0
4    0.0    0.0  0.0     0.0       0.0   0.0           0.0  0.000000   0.0

   agar  ...   wc  weekend  weti  wifi  wisata  wujudkan  yaitu  yakin  \
0   0.0  ...  0.0      0.0   0.0   0.0     0.0       0.0    0.0    0.0
1   0.0  ...  0.0      0.0   0.0   0.0     0.0       0.0    0.0    0.0
2   0.0  ...  0.0      0.0   0.0   0.0     0.0       0.0    0.0    0.0
3   0.0  ...  0.0      0.0   0.0   0.0     0.0       0.0    0.0    0.0
4   0.0  ...  0.0      0.0   0.0   0.0     0.0       0.0    0.0    0.0

       yesu  youtub
0  0.000000     0.0
1  0.169567     0.0
2  0.000000     0.0
3  0.000000     0.0
4  0.000000     0.0
```

Figure 2. TF-IDF result

Following the TF-IDF process, oversampling becomes necessary due to instances where a specific class's data quantity is significantly lower than that of other classes. The following is the difference in data before oversampling presented in Figure 3 and after oversampling presented in Figure 4, Figure 5 shows the oversampling results data.

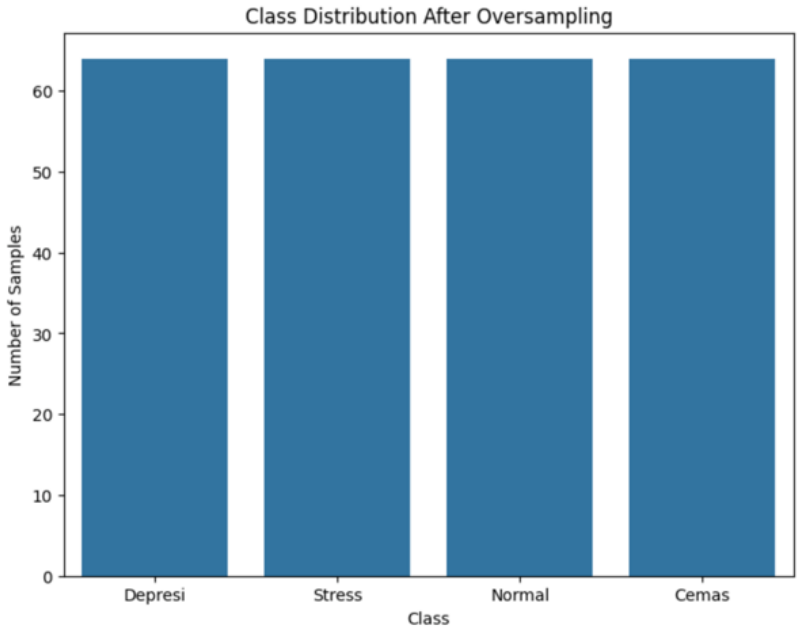

Figure 3 Data before oversampling

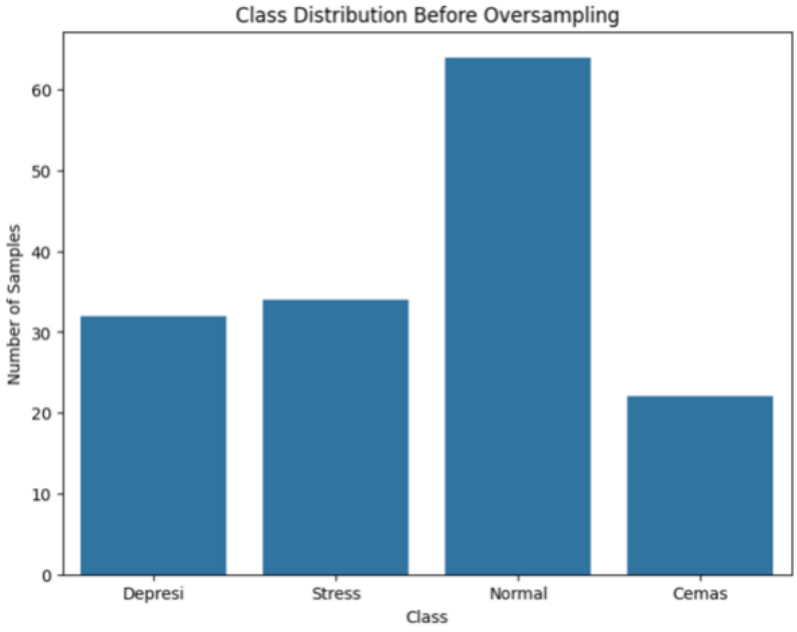

Figure 4 Data after oversampling

Data setelah Oversampling:

	abang	acara	ada	adanya	adaptasi	adik	administrasi	aduk	agak	\
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.071361	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	

	agar	...	weekend	weti	wifi	wisata	wujudkan	yaitu	yakin	yesu	\
0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
1	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.169567	
2	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
3	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
4	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	

	youtub	Label
0	0.0	Depresi
1	0.0	Stress
2	0.0	Depresi
3	0.0	Depresi
4	0.0	Normal

Figure 5. Oversampling result

After performing oversampling to balance the amount of data between the majority and minority classes, this research use cosine similarity to measure the similarity of objects in a fixed dataset. By increasing the amount of minority-class data, the model will be better at recognizing patterns in that class. Furthermore, uses cosine similarity to compare and measure the similarity between the representation vectors. This metric is crucial as it enables us to recognize relationships among more similar objects, even in the presence of class imbalance. This improves the accuracy of the model in determining the minority class after oversampling. Figure 4 presents the results of the cosine similarity calculation.

Matriks Cosine Similarity:

	0	1	2	3	4	5	6	\
0	1.000000	0.152819	0.087813	0.184315	0.242634	0.181503	0.105677	
1	0.152819	1.000000	0.088542	0.122612	0.132424	0.145739	0.104541	
2	0.087813	0.088542	1.000000	0.044562	0.036870	0.077591	0.029051	
3	0.184315	0.122612	0.044562	1.000000	0.120545	0.120525	0.106025	
4	0.242634	0.132424	0.036870	0.120545	1.000000	0.127344	0.097321	

	7	8	9	...	246	247	248	249	\
0	0.119216	0.121653	0.171879	...	0.266386	0.210518	0.259366	0.203495	
1	0.100222	0.117170	0.057952	...	0.159426	0.143614	0.207568	0.141097	
2	0.057840	0.065220	0.004881	...	0.067226	0.048811	0.107426	0.079197	
3	0.092906	0.121968	0.043239	...	0.211744	0.072271	0.234427	0.128259	
4	0.110090	0.162240	0.157030	...	0.123213	0.100788	0.213179	0.155541	

	250	251	252	253	254	255
0	0.201442	0.231771	0.227754	0.248248	0.148916	0.237009
1	0.159283	0.184620	0.119466	0.188895	0.142532	0.154707
2	0.064500	0.104120	0.072267	0.083715	0.084971	0.112288
3	0.079488	0.197101	0.135875	0.187250	0.118096	0.131945
4	0.098947	0.202811	0.224672	0.157960	0.085955	0.165144

Figure 6. Cosine similarity result

After calculating the cosine similarity to determine the similarity of objects in the dataset, the next step is to model using K-Nearest Neighbors (KNN). The cosine distance aligns the feature vectors by establishing proximity between corresponding data points, say, documents or entities. The KNN employs cosine similarity distance at a mode or modeling stage to define new data by selecting the k number of nearest neighbors that fit a classification. The larger the cosine similarity value or the smaller the distance, the higher the likelihood that the new data would inherit and share a label or category with that of the neighbors. The algorithm is highly effective in classifying or regressing tasks using the cosine similarity score, particularly when the data used has varying dimensions.

By training a database with this value of n_neighbors, KNN is derived in this study to be equal to 2 with a final accuracy figure of up to 82.69%, average precision value, recall value of 83%, and F1-Score of 82%. The most accurate label topped the performance scores, achieving the highest precision of 100 % and an F1 score of 96%. Additionally, the Depression label showed promising results, with a 93% recall rate and an 88% F1 score.

```
Model with best n_neighbours (2) produce accuracy: 0.8269

Classification Report:
              precision    recall  f1-score   support

       Cemas       1.00      0.93      0.96        14
     Depresi       0.82      0.93      0.88        15
      Normal       0.64      0.90      0.75        10
      Stress       0.88      0.54      0.67        13

    accuracy                           0.83        52
   macro avg       0.84      0.83      0.81        52
weighted avg       0.85      0.83      0.82        52
```

```
Accuracy for a given K value:
K=3: 0.8077
K=4: 0.7692
K=5: 0.7115
K=6: 0.6923
K=7: 0.6731
K=8: 0.5769
K=9: 0.5577
```

Figure 7. KNN with training data

After training the model with training data with n_neighbors = 2 and achieving 82.69% accuracy, the next step involved validating the model with validation data was carried out. At this stage, the model underwent testing using the validation data, which comprises data unseen by the model previously, to assess the model's capability to generalize the learned patterns from training. Evaluation scores of 82.35% in model performance evaluation, with an average precision value of 83%, recall of 82%, and f1-score of 80%. Anxious label achieved the highest precision rating of 100% and f1-score of 97%, followed by the Depression label with an f1-score of 82% and a recall rating of 100%.

```
Model with best n_neighbours (2) produce accuracy: 0.8431 on validation data

Classification Report (Validasi):
              precision    recall  f1-score   support

       Cemas       0.95      1.00      0.97        18
     Depresi       0.75      1.00      0.86         9
      Normal       0.80      0.86      0.83        14
      Stress       0.80      0.40      0.53        10

    accuracy                           0.84        51
   macro avg       0.82      0.81      0.80        51
weighted avg       0.84      0.84      0.83        51
```

Figure 8. Evaluation KNN model with validation data

Following the model training with validation data using the parameter n_neighbors = 2, which resulted in an accuracy of 82.35%, the subsequent step involved evaluating the model with the testing data. During this stage, the model underwent testing with a dataset entirely distinct from the training and validation data to ensure thorough evaluation against new data. The anxious label demonstrated the highest performance, attaining 100% precision, 100% recall, and 100% f1-score.

```
Accuracy on test data: 0.8654

Classification Report (Test):
              precision    recall  f1-score   support

       Cemas       1.00      1.00      1.00        10
     Depresi       0.71      0.92      0.80        13
      Normal       0.88      1.00      0.93        14
      Stress       1.00      0.60      0.75        15

    accuracy                           0.87        52
   macro avg       0.90      0.88      0.87        52
weighted avg       0.89      0.87      0.86        52
```

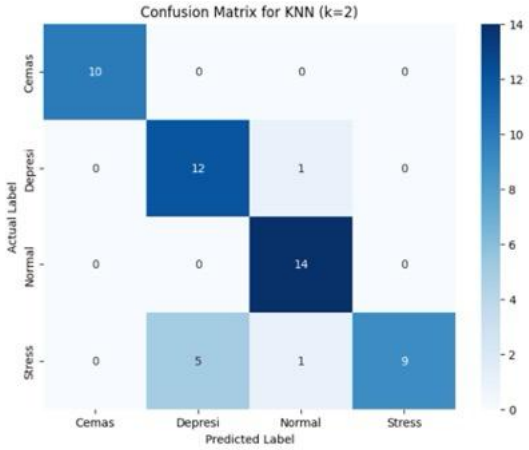

Figure 9. Evaluation KNN model with testing data

Creating a model for mental health testing is the next step. The system aims to diagnose mental health disorders in individuals using input data, such as symptoms or psychological responses. The system can use the trained and tested KNN model to classify a patient's mental health condition based on test results, surveys, and other observational data. With an accuracy of 84.62% from previous tests, the model has demonstrated the ability to make predictions with a high success rate. The initial recommendations or diagnoses provided by the system assist mental health professionals in making further decisions, enhancing the process of diagnosing and treating patients.

Figure 10. System prediction results

After the model was applied in the mental health diagnosis system and tested with new data, the KNN method model provided good accuracy of 88.46% by producing predictions of mental health depression for new cases that had been tested on the mental health diagnosis system.

DISCUSSION

Research conducted by U. Bancin, B. Bustami, and L. Rosnita [29] with the title "Expert System for Diagnosis of Mental Health Disorders in Students Using Case-Based Reasoning Method With a Web-Based Positive Psychology Approach", It found that the Case-Based Reasoning Method with the Jaccard Similarity Coefficient works by looking for similarities between new cases and old cases that are stored in the knowledge base. Problem-solving is obtained from the oldest case with the highest similarity. The knowledge base stores symptoms associated with mental disorders such as panic, anxiety, stress, and depression, which serve as the basis for the calculation of similarity. The expert systems have proven effective in applying the case-based reasoning method to determine the possibility of mental health disorders in students, based on the symptom data selected in the system.. The study tested the system using 20 test data points containing four categories of disorders and 38 symptoms, achieving an accuracy rate of 85%.

The study's results demonstrate a systematic approach to diagnosing mental health conditions using structured data comprised 238 cases with mental condition labels, which were further divided into 35 anxiety cases, 48 depression cases, 93 normal cases, and 62 stress cases. This research partitioned the data

into 80% training data, 20% test data, and used 20% of the training data as validation data. d as validation data. This study conducts several stages of the text mining process on the training data, which include data cleaning, stopword removal, normalization, and tokenization, all aimed at generating more consistent and representative data. It then continues this stage by calculating the weight of important words in the document using TF-IDF.

Then oversampling to overcome the imbalance in data between classes, enabling the model to recognize patterns in the minority class. In addition, the calculation of similarity between data using cosine similarity helps the model to recognize patterns even in unbalanced data. Subsequently, the K-Nearest Neighbors method was applied, utilizing cosine similarity to assess the proximity between data points. In this research, the optimal K value was determined to be 2, resulting in an accuracy of 82.69% on the training data, with an average precision of 85%, recall of 83%, and an F1-score of 82%. The anxious label has the highest precision, which is 100%, and F1 score of 96%, followed by the depression label with a recall of 93% and F1 score of 88%.

Validation data evaluation showed an accuracy of 82.35%, averaging a precision 83%, with a recall value of 82%, and F1 score of 80%. As stated earlier, anxiety -the label had the highest performance with a precision 100%-and an F1 score of 97%. The final evaluation of the test data yielded improved results, with an accuracy of 84.62%, an average precision of 88%, a recall of 85%, and an F1 score of 84%. The anxiety label achieved maximum performance with a precision, recall, and F1 score of 100% each.

KNN-trained and tested models like these can be utilized for mental health diagnosis systems. This system can predict the mental states of individuals based on their reported symptom data or psychological responses, thereby assisting mental health professionals in making informed decisions. This particular feature highlights the significant potential in identifying mental states using KNN-based systems.

CONCLUSION

This study successfully developed a K-Nearest Neighbors (KNN)-based model for mental health diagnosis, achieving a significant level of accuracy. It has been demonstrated that the data processing steps, involving text mining, oversampling to address imbalanced data, and cosine similarity calculations, enhance the model's performance. With an optimal value of K equal to 2, the model achieved 84.62% accuracy for the test data besides an average precision of 88% and recall of 85% and F1-score of 84%. The anxiety label demonstrated the highest attained performance with perfect precision, recall, and F1-score (100%).

These results demonstrate that the KNN model can effectively diagnose mental health conditions based on symptom data or psychological responses. The developed system would significantly aid professionals in expediting processes, enhancing diagnostic accuracy, and guiding decision-making for improved mental healthcare.

Further research can focus on utilizing other machine learning algorithms such as SVM or Deep Learning to enhance model accuracy. Expanding the model by including additional diagnostic classes (e.g., bipolar disorder or schizophrenia) can broaden the scope of diagnosis.

REFERENCES

[1] K. Amri *et al.*, "WEB-Based e-Personal Counseling (e-PC) Model Reduces Anxiety in the Face of National Examination," *KnE Social Sciences*, vol. 2023, pp. 244–253, 2023, doi: 10.18502/kss.v8i4.12905.

[2] P. Harsani, E. Erniyati, and D. Kurnia, "Strengthening School Counseling Guidance Activities through e-counseling," *International Journal of Ethno-Sciences and Education Research*, vol. 2, no. 1, pp. 43–48, 2022, doi: 10.46336/ijeer.v2i1.241.

[3] R. Rahim, W. Purba, M. Khairani, and R. Rosmawati, "Online Expert System for Diagnosis Psychological Disorders Using Case-Based Reasoning Method," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Nov. 2019. doi: 10.1088/1742-6596/1381/1/012044.

[4] K. Moon and A. Jetawat, "Predicting Lung Cancer with K-Nearest Neighbors (KNN): A Computational Approach," *Indian J Sci Technol*, vol. 17, no. 21, pp. 2199–2206, 2024, doi: 10.17485/IJST/v17i21.1192.

[5] A. Ramadhan, L. Lindawati, and M. M. Rose, "Komparasi Algoritma Neural Network dan K-Nearest Neighbor Dalam Mendeteksi Malware Android," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3538.

[6] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019*, Institute of Electrical and Electronics Engineers Inc., May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[7] W. A. Ridho, T. Wuryandari, and A. R. Hakim, "PERBANDINGAN KINERJA METODE KLASIFIKASI K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINES PADA DATASET PARKINSON," *Jurnal Gaussian*, vol. 12, no. 3, pp. 372–381, Feb. 2024, doi: 10.14710/j.gauss.12.3.372-381.

[8] X. Tian, "Construction of mental health education system for college counselors based on web technology," *IOP Conf Ser Mater Sci Eng*, vol. 750, no. 1, 2020, doi: 10.1088/1757-899X/750/1/012001.

[9] S. Mulyana, S. Hartati, and R. Wardoyo, "Case Based Reasoning for Diagnosing Types of Mental Disorders and Their Treatments," in *Communications in Computer and Information Science*, 2019, pp. 325–335. doi: 10.1007/978-981-15-0399-3_26.

[10] A. Tyagi, V. P. Singh, and M. M. Gore, "An Efficient Automated Detection of Schizophrenia Using k-NN and Bag of Words Features," *SN Comput Sci*, vol. 4, no. 5, 2023, doi: 10.1007/s42979-023-01947-2.

[11] H. Benfriha *et al.*, "A new approach for case acquisition in CBR based on multi-label text categorization: a case study in child's traumatic brain injuries," *International Journal of Computing and Digital Systems*, vol. 10, no. 1, pp. 1003–1018, 2021, doi: 10.12785/IJCDS/100190.

[12] A. Leśniak and K. Zima, "Cost calculation of construction projects including sustainability factors using the Case Based Reasoning (CBR) method," *Sustainability (Switzerland)*, vol. 10, no. 5, 2018, doi: 10.3390/su10051608.

[13] T. M. R. Gunung, S. S. Lubis, M. Siregar, P. J. N. Simanjuntak, and A. Jinan, "IMPLEMENTASI METODE CASED BASED REASONING (CBR) DALAM SISTEM PAKAR UNTUK MENDAPATKAN DIAGNOSIS ANXIETY DISORDERS," *Jurnal Teknologi Terpadu*, vol. 10, no. 2, 2024.

[14] Y. Ronaldo Juliantino and E. Rosana Widasari, "Implementasi Metode K-Nearest Neighbor untuk Deteksi Dini Stres Berbasis Neurosky Electroencephalogram Sensor," 2025. [Online]. Available: http://j-ptiik.ub.ac.id

[15] H. A. Musril, S. Saludin, W. Firdaus, S. Usanto, K. Kundori, and R. Rahim, "Using k-NN Artificial Intelligence for Predictive Maintenance in Facility Management," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 10, no. 6, pp. 1–8, 2023, doi: 10.14445/23488379/IJEEE-V10I6P101.

[16] A. Fahmi Limas, R. Rosnelly, and A. Nursie, "A Comparative Analysis on the Evaluation of KNN and SVM Algorithms in the Classification of Diabetes," *Scientific Journal of Informatics*, vol. 10, no. 3, p. 251, 2023, doi: 10.15294/sji.v10i3.44269.

[17] A. Pamuji, "Performance of the K-Nearest Neighbors Method on Analysis of Social Media Sentiment," *JUISI*, vol. 07, no. 01, 2021.

[18] A. Masitha, M. K. Biddinika, and H. Herman, "K Value Effect on Accuracy Using the K-NN for Heart Failure Dataset," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 3, pp. 593–604, Jul. 2023, doi: 10.30812/matrik.v22i3.2984.

[19] N. A. Saputra, K. Aeni, and N. M. Saraswati, "Indonesian Hate Speech Text Classification Using Improved K-Nearest Neighbor with TF-IDF-ICSρF," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 21–30, Feb. 2024, doi: 10.15294/sji.v11i1.48085.

[20] A. Masitha and M. Kunta Biddinika, "KLIK: Kajian Ilmiah Informatika dan Komputer Preparing Dual Data Normalization for KNN Classfication in Prediction of Heart Failure," *Media Online*, vol. 4, no. 3, pp. 1227–1234, 2023, doi: 10.30865/klik.v4i3.1382.

[21] A. Lubis, "Leveraging K-Nearest Neighbors with SMOTE and Boosting Techniques for Data Imbalance and Accuracy Improvement," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1625–1638, Dec. 2024, doi: 10.47738/jads.v5i4.343.

[22] D. Syarif Sihabudin Sahid, "Jurnal Politeknik Caltex Riau Using KNN Algorithms for Determining the Recipient of Smart Indonesia Scholarship Program," 2021. [Online]. Available: https://jurnal.pcr.ac.id/index.php/jkt/

[23] J. Elektronik, I. K. Udayana, B. E. Sutrisna, and E. Karyawati, "Comparison of K-Nearest Neighbor And Modified K-Nearest Neighbor With Feature Selection Mutual Information And Gini Index In Informatics Journal Classsification," *Jurnal Elektronik Ilmu Komputer Udayana*, vol. 10, no. 3, Feb. 2022, [Online]. Available: https://www.neliti.com/id/conferences/semnasif

[24] K. Hadi and E. Utami, "Analysis of K-NN with the Integration of Bag of Words, TF-IDF, and N-Grams for Hate Speech Classification on Twitter," 2024.

[25] N. Nasution, F. Feldiansyah, A. Zamsuri, and M. A. Hasan, "Synthetic Minority Oversampling Technique for Efforts to Improve Imbalanced Data in Classification of Lettuce Plant Diseases," *JURNAL TEKNOLOGI DAN OPEN SOURCE*, pp. 31–40, Feb. 2023, doi: 10.36378/jtos.v6i1.2883.

[26] O. N. Oyelade and A. E. Ezugwu, "A case-based reasoning framework for early detection and diagnosis of novel coronavirus," *Inform Med Unlocked*, vol. 20, no. July, p. 100395, 2020, doi: 10.1016/j.imu.2020.100395.

[27] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci Rep*, vol. 12, no. 1, pp. 1–12, 2022, doi: 10.1038/s41598-022-10358-x.

[28] M. Danny, A. Muhidin, and A. Jamal, "Application of the K-Nearest Neighbor Machine Learning Algorithm to Preduct Sales of Best-Selling Products," *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 1, pp. 255–264, Jun. 2024, doi: 10.47709/brilliance.v4i1.4063.

[29] U. Bancin, B. Bustami, and L. Rosnita, "Expert System For Diagnosis of Mental Health Disorders in Students Using Case-Based Reasoning Method With a Web-Based Positive Psychology Approach," *International Journal of Engineering, Science and Information Technology*, vol. 4, no. 4, pp. 135–143, Oct. 2024, doi: 10.52088/ijesty.v4i4.592.

.