Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

# SVM-Based Pediatric Disease Classification Model from the Balinese Lontar Usada Rare Manuscript

**I Gusti Made Ngurah Ari Bhawanaputra[1)], I Gede Iwan Sudipa[2)*], Ni Putu Suci Meinarni[3)], I Gusti Ayu Agung Mas Aristamy[4)], Indra Pratistha[5)]**

[1,2*,3,4,5)]Fakultas Teknologi dan Informatika, Program Studi Informatika, Institut Bisnis dan Teknologi Indonesia, Bali, Indonesia

[1)]aribhawana012@gmail.com, [2)*]iwansudipa@instiki.ac.id, [3)]sucimeinarni@instiki.ac.id, [4)]agungmas.aristamy@instiki.ac.id, [5)]indra.pratistha@instiki.ac.id

**Submitted** :Oct 29, 2025 | **Accepted** : Dec 16, 2025 | **Published** : Jan 04, 2026

**Abstract:** Lontar Usada Rare is a traditional Balinese manuscript containing pediatric medical knowledge based on local wisdom, yet its narrative format limits accessibility and utilization in modern contexts, while its physical fragility threatens long-term preservation. This study aims to develop a pediatric disease classification model using a Support Vector Machine (SVM) combined with Term Frequency–Inverse Document Frequency (TF-IDF) weighting to support the digitalization of Balinese traditional medicine. A total of 422 data samples were collected through expert interviews and manuscript analysis, covering symptoms, disease types, herbal ingredients, and treatment procedures. The research stages included text preprocessing (cleansing, tokenizing, stopword removal, stemming), manual labeling into 35 disease classes, and model evaluation using five train–test split ratios (80:20 to 60:40) with variations of the complexity parameter C (0.5, 1, 10, 100, 1000). The best performance was achieved using C=10 with an 80:20 ratio, resulting in 87.06% accuracy, 91.55% precision, 87.06% recall, and an F1-score of 87.96%. Confusion matrix analysis showed strong classification performance for most classes, although minority classes with overlapping symptoms exhibited misclassification. Overall, the TF-IDF and linear SVM combination effectively classifies pediatric disease symptoms from Lontar Usada Rare and contributes to the preservation and digital transformation of Balinese traditional medical knowledge for potential modern healthcare applications.

**Keywords:** Classification, Lontar Usada Rare, Balinese Traditional Medicine, Support Vector Machine, TF-IDF

## INTRODUCTION

Bali is known as a region rich in tradition and culture, including local knowledge passed down through generations through ancient manuscripts in the form of lontar. One of the most important is Lontar Usada, which contains explanations about the philosophy, diagnosis, and procedures of traditional Balinese medicine (Adnyana, 2020) . Of the various types of lontar, Usada Rare specifically discusses treatment for children, with information on disease symptoms, herbs, and local wisdom-based healing methods. Various medicinal plants mentioned, such as turmeric, temu ireng, and ginger, are also proven to have biological properties in relieving inflammation and increasing endurance (Adnyana, 2021) . Despite its importance, lontar manuscripts face significant preservation challenges in the modern era. Many manuscripts have deteriorated physically, with faded writing and fragile sheets. Some are in danger of being lost due to improper storage. Digitization efforts have been made but remain limited, manual, and not thoroughly coordinated (Balipost.com, 2024). The preservation of ancient manuscripts through digitization is crucial, as affirmed by BRIN and the Lontar Study Center of Udayana University. Digitization can save lontar from physical damage while expanding public access(Brin.go.id, 2025; Goodnewsfromindonesia.id, 2025). The Indonesian government's commitment has also been outlined in regulations, such as Law Number 5 of 2017 concerning the Promotion of Culture and Minister of Health Regulation 37 on traditional health services based on local wisdom (Peraturan Menteri Kesehatan Nomor 37 Tahun 2017 Tentang Pelayanan Kesehatan Tradisional Integrasi, 2017; UU Nomor 5 Tahun 2017 Tentang Pemajuan Kebudayaan, 2017). However, Lontar Usada Rare remains accessible only in a few institutions such as Udayana University and Gedong Kirtya Singaraja, necessitating new strategies to overcome this limitation.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

With the development of information technology, digitization of cultural heritage serves both preservation and wider utilization for education, research, and healthcare applications. Information in Lontar Usada Rare is presented in long narrative form, requiring classification to organize it into a more structured format. This includes grouping disease symptoms, disease types, and treatment methods. Text classification techniques can bridge traditional knowledge with digital technology, so that information that was previously difficult to access can be presented more systematically and efficiently. n this research, the Natural Language Processing (NLP)-based Support Vector Machine (SVM) method is employed due to its proven effectiveness in handling large and unbalanced text datasets. A number of previous studies have also proven the effectiveness of SVM, for example with 85% accuracy in credit data classification (Putra Asana & Della Tirta Yanti, 2023), 99.88% in colored texture classification using global-local feature extraction (Navarro & Perez, 2019) , and demonstrated robust performance in polymer classification tasks with high-dimensional datasets (Malashin et al., 2025). These findings show that SVM is able to provide accurate classification results for complex and unstructured data.

Unlike previous SVM-based medical text classification studies that focus on standardized clinical data, this research introduces a cultural-linguistic adaptation pipeline for traditional Balinese medical texts. The novelty lies in three key aspects: (1) domain adaptation that transforms archaic Balinese medical vocabulary into modern NLP-compatible representations through customized normalization and stemming processes tailored for Old Balinese texts, (2) semantic-based disease grouping to address class imbalance by clustering diseases with similar symptomatology into unified categories (e.g., multiple Tiwang variants consolidated into a single "Tiwang" class based on shared clinical manifestations), and (3) semantic mapping between narrative-style symptom descriptions characteristic of traditional manuscripts and structured disease classification categories, enabling pattern recognition despite linguistic ambiguity inherent in historical medical texts (Sun et al., 2017; Venkataraman et al., 2020). This approach demonstrates that SVM, when combined with culturally-adapted preprocessing techniques, effectively processes historical medical manuscripts characterized by linguistic diversity, non-standardized spelling, and sparse textual representation, thereby establishing a methodological framework applicable to similar cultural heritage digitization projects.

Therefore, this research aims to develop a classification model for childhood diseases based on Lontar Usada Rare with the SVM approach, as an effort to support the digitization of Balinese cultural heritage while expanding its use in the field of traditional health in the digital era.

## LITERATURE REVIEW

Data classification, particularly on text and health data, has been the focus of many studies utilizing machine learning algorithms. One of the most commonly used methods is the Support Vector Machine (SVM) due to its ability to handle both structured and unstructured data, especially when combined with Natural Language Processing (NLP) techniques. SVM has proven to be effective in producing optimal class separation through the use of maximum margins, so it is widely adopted in text-based classification research and medical data. A number of previous studies support the superiority of SVM in various contexts. (Hidayat et al., 2024) , for example, successfully classified text related to Sustainable Development Goals (SDGs) using multiclass SVM with Term Frequency-Inverse Document Frequency (TF-IDF) word representation, and recorded high accuracy of up to 98.08%. Similar success was also shown by (Dewanti et al., 2025) in the classification of stunting disease in infants, where the Radial Basis Function (RBF) kernel was used to overcome data imbalance and resulted in an accuracy of 95.26%. Another study by (Rahayu & Yamasari, 2024) on stroke disease classification showed that the polynomial kernel is superior to other kernels with an accuracy of 78.86%. Meanwhile (Ropikoh et al., 2021) , proves the effectiveness of SVM in detecting Covid-19 hoax news, with accuracy reaching 97.06% on linear kernels. Even in the realm of sentiment analysis, (Verdikha & Yulianto, 2025) used a combination of Word2Vec and SVM to manage SIREKAP application review data, resulting in 95% accuracy with high precision, recall, and F1-score.

Despite these advances, existing studies predominantly focus on modern clinical text with standardized terminology and consistent orthography. A significant research gap exists in processing historical medical manuscripts characterized by archaic vocabulary, dialectal variations, and narrative-based symptom descriptions. Recent work on automatic pathology report classification (Fiebig et al., 2018), clinical text classification with weak supervision (Kim et al., 2021), medical subdomain classification (Weng et al., 2017), and class imbalance handling (Lu et al., 2022; Zhu et al., 2024) has demonstrated promise in medical domains. However, these approaches assume standardized clinical terminology and structured data formats, which are absent in pre-modern texts such as Lontar Usada Rare. Furthermore, no studies have addressed the challenge of semantic mapping between narrative-style symptom descriptions and structured disease categories in traditional medical contexts.

This research addresses these gaps by introducing a culturally-adapted NLP pipeline specifically designed for Old Balinese medical texts. The novelty lies in three key contributions: (1) customized normalization and stemming processes that handle archaic vocabulary and non-standardized spelling variations unique to lontar manuscripts, (2) semantic-based disease clustering to address severe class imbalance by grouping diseases with

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

similar symptomatology, and (3) domain adaptation techniques that transform narrative symptom descriptions into structured features suitable for machine learning classification. Using TF-IDF for feature extraction and evaluating model performance through confusion matrix metrics (accuracy, precision, recall, and F1-score), this study establishes a replicable framework for digitizing traditional medical knowledge from cultural heritage texts.

## METHOD

This research includes a series of stages that are systematically organized to produce a pediatric disease classification model based on the traditional Balinese text Lontar Usada Rare. Each stage is designed to ensure data quality, method accuracy, and overall model performance evaluation. The research data sources were obtained through two methods, namely interviews with traditional Balinese medicine experts and the Lontar Usada Rare manuscript from Gedong Kirtya Singaraja which has been translated into Indonesian. From these two sources, a total of 422 data were collected, containing information on symptoms of diseases, types of diseases, medicinal ingredients, and treatment procedures. The research stages were organized as shown in Fig. 1. Research Flow.

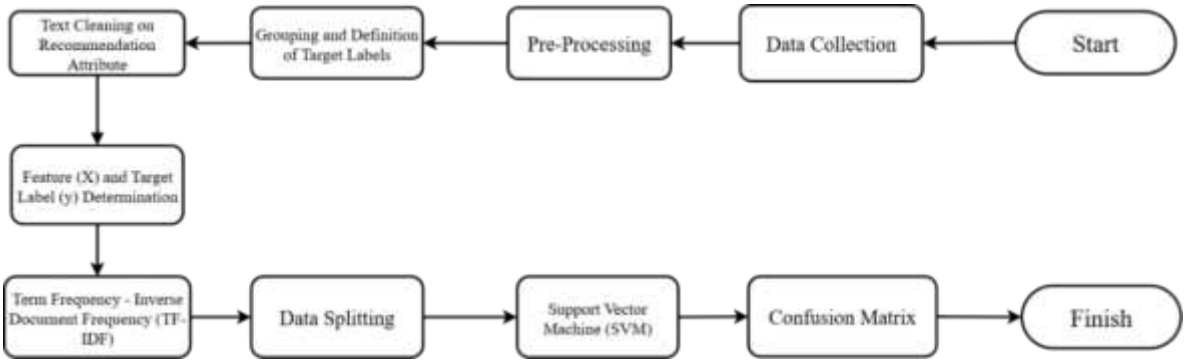


Fig. 1. Research flow

The first step is Data Collection, which is the process of collecting and organizing data from literature sources and interview results to be ready for use in the modeling stage. The next stage is Pre-Processing, which serves to prepare text data to be ready for processing by the model. This process includes cleansing (removing irrelevant characters), case folding (uniforming lowercase letters), normalization (uniforming words), tokenizing (breaking text into words), stopword removal (removing words that are not important), and stemming (returning words to their basic form). This stage is important to reduce noise and ensure data consistency (Dag, 2020; Pradana & Hayaty, 2019) .

The next stage is Grouping and Definition of Target Labels, where data is grouped into 35 disease classes based on symptom similarity. After that, Text Cleaning on Recommendation Attribute is performed, which aims to clean the text on the treatment recommendation attribute to make it more consistent and easy to process by the system. The next step is Feature (X) and Target Label (y) Determination, where symptom attributes are used as features (X) and disease types as target labels (y). Because the label is still categorical, a label encoding process is carried out so that the data can be converted into a numerical form that can be processed by the algorithm. After that, the Term Frequency-Inverse Document Frequency (TF-IDF) method is applied, which is used to give weight to each word based on its importance in the document. This process helps identify the most relevant and informative words.

The next stage is Data Splitting, where the dataset is divided into training data and test data using five different ratios, namely 80:20, 75:25, 70:30, 65:35, and 60:40. The purpose of this test is to find the most optimal data split ratio before the model is applied (Hidayat et al., 2024) . This approach of testing multiple ratios is in line with the findings of (Abdalla, 2022) , which shows that the optimal data sharing ratio may vary depending on the characteristics of the dataset, and experimentation to determine the best ratio is an important practice to ensure the generalizability of the model. The core stage of the research is the application of the Support Vector Machine (SVM) algorithm with a linear kernel, which is effective in handling high-dimensional data and producing stable classifications. As the final stage, the model evaluation was carried out using Confusion Matrix to calculate the accuracy, precision, recall, and F1-score values, so as to thoroughly describe the performance of the model in classifying diseases based on symptoms (Dewanti et al., 2025) . The use of these metrics is in line with the standard practice of model evaluation, where according to (Grandini et al., 2020) , the combination of accuracy, precision, recall, and F1-score derived from Confusion Matrix is essential to provide a holistic and reliable assessment of model performance, especially in medical classification tasks.

### Data Collection

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data collection in this research was conducted through two main sources, namely primary and secondary data. Primary data was obtained through an interview with Dr. Nyoman Sridana, S.Kes.H., M.Si., owner of PT Vision Bali Herbal Production Usada Taru Pramana, who has extensive experience in the production of traditional Balinese herbal medicines since 2013. This interview provided in-depth information on the types of diseases, symptoms, and herbal treatment methods listed in Lontar Usada Rare. Meanwhile, secondary data was obtained from a copy of the Lontar Usada Rare that has been booked and translated into Indonesian, which was collected directly from Gedong Kirtya Singaraja. The data totaled 422 items and contained various types of diseases, accompanying symptoms, herbal medicines, and traditional treatment procedures. This data source served as the basis for analysis and classification in the study.

Table. 1. Lontar Usada Rare dataset

| Lontar Usada Rare Data |
|---|
| Jika ada bayi yang sakit, jika terlihat pada putih matanya biru, tangan dan kakinya dingin, tangisnya agak serak dan merdupkan mata, sakit bayi ini, penyakitnya pada tulang ubun-ubun renggang disebut belahan. Bahan obatnya sembung kedis, ambil akarnya, beras merah, lempuyang, mesui, Semua lumatkan. Tempelkan pada ubun-ubunnya |
| Jika seperti agak putih matanya si bayi, serta uratnya seperti berdarah, tenaganya tidak stabil, bibirnya keropos, panas badan si bayi itu, penyakitnya disebut guaman. Sarana : biji mentimun, buah paspasan, subatah enau, wong papah idung, semua dikuskus, lumatkan pada mulutnya. |
| Jika seperti ada unged pada tubuh menyebar, sikap kaki dan tangan ditelungkupkan , mata kemerahan , disebut tiwang penyu. Sarana obatnya : babakan ceremai, pala, menyan, 484unged, sinrong, gegambiran, arak, lulurkan |
| Putih matanya biru, tangan dan kakinya dingin, tangisnya agak serak dan merdupkan mata, Penyakitnya pada tulang ubun-ubun renggang disebut belahan, Bahan obatnya sembung kedis, ambil akarnya, beras merah, lempuyang, mesui, Semua lumatkan. Tempelkan pada ubun-ubunnya. |
| Agak putih matanya si bayi, serta uratnya seperti berdarah, tenaganya tidak stabil, bibirnya keropos, panas badan si bayi itu, Penyakitnya disebut guaman, Biji mentimun, buah paspasan, subatah enau, wong papah idung, Semua dikuskus, lumatkan pada mulutnya. |
| Hidung bayi tersumbat dan berair, Padang lepas, daging kemiri, pulasai, lengkuas 3 iris, Sembar pada taneng, tengah-tengah |
| Telinga bayi terlihat seperti mengeluarkan air, Sirih lanang, minyak wijen, Dipipis lalu ditiupkan pada telinganya. |

**Meta Data**

After the data was collected, the next step was the preparation of metadata to provide a more systematic structure to the information contained in Lontar Usada Rare. The metadata is organized into five main columns, namely Symptoms of Illness, Type of Illness, Class of Illness, Medicinal Ingredients, and Treatment Procedures.

Table. 2. Metadata of the Lontar Usada Rare Dataset

| No | Symptom of Illness | Type of Illness | Illness Type Class | Medicine Materials | Treatment Procedure |
|---|---|---|---|---|---|
| 1. | Putih matanya biru, tangan dan kakinya dingin, tangisnya agak serak dan merdupkan mata | Penyakitnya pada tulang ubun-ubun renggang disebut belahan. | Belahan | Bahan obatnya sembung kedis, ambil akarnya, beras merah, lempuyang, mesui | Semua lumatkan. tempelkan pada ubun-ubunnya |
| 2. | Agak putih matanya si bayi, serta uratnya seperti berdarah, tenaganya tidak stabil, bibirnya keropos, panas badan si bayi itu | Penyakitnya disebut guaman | Guaman | Biji mentimun, buah paspasan, subatah enau, wong papah idung | Semua dikuskus, lumatkan pada mulutnya |
| 3. | Agak kemerahan tubuh sang bayi | Disebut tiwang brahma | Tiwang | Daun dapdap yang muda, empol ending merah, triketuka | Lulurkan |
| ..... | | | | | |
| 100. | Muncul bercak merah pada tubuh si bayi | Gatal | Gatal-gatal | Daun dapdap wong yang sedang, kapur tohor | Dilulurkan |
| 101. | Bayi batuk-batuk dan suhu badan panas dingin | Demam | Panas dingin | Daun kemuning, sulasih harum, kencur, temu tis, kelapa bakar | Disembar |
| ..... | | | | | |
| 422. | Jika matanya terlihat biru dan putih, tangan serta kaki terasa dingin, menangis dengan suara agak serak, serta tubuh meredup dan melemas, dan terdapat belahan di ubun-ubun. | Belahan pada ubun ubun | Belahan | Sembung kedis lulurnya, beras merah, lempuyang, mesui, semua lumatkan. | Di tempel pada ubun ubunnya. |

The process of classifying classes was done manually by combining diseases that had similar symptoms or characteristics, for example, several variants of Tiwang disease were combined into one class. The creation of

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

metadata not only facilitates the classification process, but also ensures that the data can be analyzed in a more targeted manner and provides a clear representation of the relationship between symptoms, disease types, herbs, and traditional Balinese medicine methods.

**Pre-processing**

Pre-processing is an important initial stage in Natural Language Processing (NLP) that aims to prepare raw text data for effective analysis. This stage includes several sequential steps, namely cleansing to remove irrelevant elements such as punctuation marks, numbers, and symbols (Kunilovskaya, 2021) , case folding to uniform letters to be small so that capitalization variations do not affect the analysis results (Wabula et al., 2023) , and normalization to unify writing variations or non-standard words into standard forms (Saputra et al., 2024) . Furthermore, tokenization is carried out to break the text into word units (Chen et al., 2022; Gastaldi et al., 2025) , stopword removal to remove common words that do not make an important contribution to the meaning of the text (Sarica & Luo, 2021) , and stemming to return words to their basic form so that word variations such as "running" and "running around" are treated the same. The systematic application of these preprocessing stages is proven to improve data quality and the effectiveness of text classification models, as shown in various previous studies (Dag, 2020; Khan et al., 2024; Pradana & Hayaty, 2019).

**Grouping and Definition of Target Labels**

After the pre-processing stage is complete, this research continues with the process of grouping and definition of target labels. Given the wide variety of disease types in Lontar Usada Rare and the uneven distribution of the data, a manual clustering strategy was carried out by combining diseases that have similar symptoms into a more general class. For example, several variants of Tiwang disease such as "Tiwang Bangke", "Tiwang Asu", and "Tiwang Penyu" were combined into the class "Tiwang", while the symptoms "Heat", "Fever", and "Cough" were placed in the same group based on similar characteristics. This clustering resulted in a total of 35 disease classes used as classification targets. This approach aims to simplify the label space, reduce the imbalance of data distribution, and help the model recognize patterns more stably and representatively.

To facilitate processing by the algorithm, each disease class that has been formed is then converted into numerical form through the label encoding process. This implementation was done using Pandas and LabelEncoder in Python, where each disease type was given a certain numerical identity, such as "Barah" with label 0, "Curek" with label 1, "Diarrhea" with label 2, to "Reddish body" with label 34. The transformation of data from text to numerical labels allows the Support Vector Machine (SVM) model to read and process data more efficiently. Thus, this labeling process serves as a bridge between the traditional linguistic representation of lontar and the technical requirements of machine learning-based modeling, thereby improving the accuracy and consistency of classification results.

Table. 3. Disease Type Labeling

| Disease Type Class | Encoded Label | Number of Samples |
|---|---|---|
| Barah | 0 | 9 |
| Tiwang | 22 | 52 |
| Jampi | 4 | 29 |
| Upas | 23 | 14 |
| Belahan | 25 | 7 |
| Gatal-gatal | 26 | 7 |
| Batuk | 24 | 7 |
| Curek | 1 | 6 |
| Diare | 2 | 26 |
| Gangguan pernafasan | 3 | 13 |
| Kurang gizi | 5 | 3 |
| Menangis terus menerus | 6 | 8 |
| Mimisan | 7 | 6 |
| Mual-mual | 8 | 10 |
| Muntah | 9 | 11 |
| Panas | 10 | 23 |
| Panas dalam | 11 | 33 |
| Pejen | 12 | 15 |

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

| | | |
|---|---|---|
| Penguci bolong | 13 | 3 |
| Perut kembung | 14 | 27 |
| Perut melilit | 15 | 5 |
| Perut kaku | 16 | 8 |
| Sakit mata | 17 | 9 |
| Sakit perut | 18 | 9 |
| Sebaha | 19 | 15 |
| Sembelit | 20 | 8 |
| Siksik | 21 | 10 |
| Guaman | 27 | 9 |
| Inja | 28 | 13 |
| Panas dingin | 29 | 8 |
| Sakit kemaluan | 30 | 8 |
| Sakit kuning | 31 | 2 |
| Sakit telinga | 32 | 2 |
| Sula | 33 | 3 |
| Tubuh kemerahan | 34 | 2 |

Table. 3 presents an example of the results of the process of labeling the disease types contained in Lontar Usada Rare after manual clustering and conversion to numeric form using label encoding. Each entry in the table shows the name of the traditional disease class and the numerical label assigned, for example "Barah" is labeled 0, "Tiwang" is labeled 22, and "Jampi" is labeled 4. This process allows the classification model to recognize each disease type in a consistent numerical representation, making it easier for the Support Vector Machine (SVM) algorithm to process. The example in this table illustrates how the data from traditional text was transformed into a more structured format, which was then ready for use in the training and testing stages of the classification model.

**Text Cleaning on Recommendation Attribute**

At this stage, text cleaning is carried out on the recommendation attribute, especially the Medication Ingredients and Treatment Procedure columns. The process includes case folding to homogenize letters, removal of double spaces, and removal of punctuation marks to make the data more consistent and easy to read. This step was implemented using Python's clean_text_basic function which automatically simplifies the text and generates a new column of clean data. The pre-processing results show a comparison of before and after cleaning, showing improved readability and consistency of the data. This process is important for maintaining the accuracy of the treatment recommendation information as well as preparing the data to be used as support for disease classification results.

**Feature (X) and Target Label (y) Determination**

The determination of features (X) and target labels (y) is an important stage in the process of classifying childhood diseases using the Support Vector Machine (SVM) method. Features (X) represent symptom attributes obtained from the dataset by selecting relevant columns and removing label columns to leave only input variables for the model. This stage helps SVM recognize the relationship patterns between symptoms and disease types systematically. In line with (Fan et al., 2023) , the selection of features relevant to the target label plays an important role in improving the efficiency and stability of classification results. The target label (y) serves as an output that indicates the disease category such as Thywang, Fever, or Cough, which is then converted to numerical form through the label encoding process in order to be mathematically processed. In summary, the separation of features and labels can be expressed as:

$$X = [x_{ij}]_{n \times m} \quad 1) \text{ and } y = [y_i]_{n \times 1} \quad 2)$$

Where the model learns the mapping function$f(x_i) \rightarrow y_i$ to classify symptoms into corresponding disease categories. This approach ensures the data used matches the modeling needs and improves the accuracy and generalization ability of the model.

**Term Frequency-Inverse Document Frequency (TF-IDF)**

The Term Frequency-Inverse Document Frequency (TF-IDF) method is used to convert disease symptom text into a numerical representation that can be processed by classification algorithms. TF-IDF consists of two main

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

components: Term Frequency (TF), which measures how often a word appears in a document, and Inverse Document Frequency (IDF), which assesses the rarity of a word across documents. The combination of the two results in a weight that accentuates specific words and reduces the influence of common words. Mathematically, the calculation of TF, IDF, and TF-IDF are formulated as follows:

$$\mathrm{TF}(t,d) = \frac{\text{Number of occurrences of word t in document d}}{\text{Total number of words in document d}} \quad 3), \mathrm{IDF}(t) = \log\left(\frac{N}{\mathrm{DF}(t)}\right) \quad 4), \mathrm{TF-IDF}(t,f) = \mathrm{TF}(t,f) \times \mathrm{IDF}(t) \quad 5)$$

where $f(t,d)$ is the frequency of word t in document d, $|d|$ the total words in the document, N the total number of documents, and $\mathrm{DF}(t)$ the number of documents containing word t.

In this study, TF-IDF is used to extract the most relevant words from the text of pediatric disease symptoms in Lontar Usada Rare. According to (Sheridan et al., 2025) , the TF-IDF method is effective in emphasizing informative words as well as improving the performance of machine learning algorithms in text analysis. This approach allows Support Vector Machine (SVM) to focus classification on the most meaningful terms, thus improving accuracy in pediatric disease symptom pattern recognition.

As a result of the weighting process described earlier, Table 4 below presents an example of the TF-IDF weight values generated from the disease symptom data in the Lontar Usada Rare manuscript.

Table. 4. The Result of TF-IDF Weighting from Disease Symptom Texts

| Term | Document 1 | Document 2 | Document 3 |
|---|---|---|---|
| Bayi | 0,1932 | 0,1633 | 0,1471 |
| Batuk | 0,3519 | 0,2912 | 0,3096 |
| Panas | 0,2595 | 0,1690 | 0,2821 |
| Muntah | 0,1357 | 0,1072 | 0,1595 |
| Demam | 0,1357 | 0,2894 | 0,3905 |

Table. 4 displays the results of word weighting using the Term Frequency-Inverse Document Frequency (TF-IDF) method on the text of pediatric disease symptoms. The TF-IDF value represents the importance of a word in the whole set of documents by considering the frequency of its occurrence in each text. The results show that the words "cough" and "fever" have the highest weights (0.3519 and 0.3905), signifying their role as leading indicators in the context of childhood illnesses. Meanwhile, the word "heat" had a medium weight (around 0.26-0.28), while "baby" and "vomit" had lower weights (0.13-0.19), indicating their more general nature. These weighting values are then converted into numerical vectors and used by the Support Vector Machine (SVM) algorithm to quantitatively recognize patterns of relationships between words, so that the model can distinguish disease types more accurately and focus on the most relevant terms in the context of pediatric disease symptoms in Lontar Usada Rare.

**Data Splitting**

The data splitting stage is an important process in building a Support Vector Machine (SVM) based classification model, where the dataset is divided into training data to build the model and test data to evaluate its performance against new data. This study used 422 samples from 35 classes of pediatric diseases based on the Lontar Usada Rare manuscript, with data distribution using a stratified sampling method to keep the proportion of each class balanced. This approach is in line with the findings of (Nguyen et al., 2021) which shows that variations in training and testing ratios have a significant effect on model performance, and (Rácz et al., 2021) which emphasizes the importance of choosing the right ratio to achieve stable and reliable model performance.

Table. 5. Dataset Distribution in Five Plitting Scenarios

| Data Split Ratio | Training Data | Testing Data | Total Data |
|---|---|---|---|
| 80:20 | 337 | 85 | 422 |
| 75:25 | 316 | 106 | 422 |
| 70:30 | 295 | 127 | 422 |
| 65:35 | 274 | 148 | 422 |
| 60:40 | 253 | 169 | 422 |

The results of data sharing as shown in Table. 5 are the basis for evaluating the effect of ratios on four main metrics, namely accuracy, precision, recall, and F1-score. Through five test scenarios, we can assess the balance between the model's ability to learn patterns (learning capacity) and generalize new data (generalization ability). This approach also serves as an internal validation of the stability of the SVM algorithm, ensuring that the best ratio chosen is a consistent and reliable result. Thus, the data splitting stage is not only a technical process, but

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

also a strategic step in ensuring the quality and credibility of the Lontar Usada Rare-based pediatric disease classification results.

**Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a widely used machine learning algorithm for classification and regression by finding the best hyperplane that separates data between classes with maximum margin. The closest data points of the hyperplane, called support vectors, play an important role in determining the decision boundary, resulting in a model that is stable and has high generalizability to new data (Guido et al., 2024) . The main advantage of SVM is its ability to handle data that is not linearly separable through the use of kernel tricks, which map the data to a higher dimensional space in order to be optimally separated. In this study, a linear kernel is used because it matches the characteristics of text data in the form of high-dimensional vectors and is sparse (Amaya-Tejera et al., 2024) Mathematically, the SVM classification function can be expressed as:

$$f(x) = w.x + b \quad 6) \quad \text{or in kernel function form } f(x) = \sum_{i=1}^{m} a_i\, y_i K(x.xi) + b \quad 7)$$

where$\alpha_i$ is the model parameter,$y_i$ the class label,$K(x, xi)$ the kernel function, and b is the bias.

In this study, the Support Vector Machine (SVM) algorithm is used to classify text from Lontar Usada Rare which includes disease symptoms, disease types, herbal medicine ingredients, and treatment procedures. The best parameter selection was done based on the highest accuracy value from the test results, then the model was evaluated using precision, recall, and F1-score metrics to comprehensively assess its performance. This approach produces a model that is not only accurate, but also has good generalization ability to new data.

**Confusion Matrix**

Confusion Matrix is an evaluation tool used to measure the performance of classification models by comparing predicted results to actual labels. This matrix consists of four main components, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These four values become the basis for calculating evaluation metrics such as accuracy, precision, recall, and F1-score, which describe the level of accuracy and ability of the model to classify data (Sathyanarayanan & Roopashri Tantri, 2024) . Accuracy indicates the proportion of correct predictions from the overall data, precision assesses how relevant the positive predictions produced by the model are, recall measures the ability of the model to recognize all positive data. F1-score is a harmonic mean between precision and recall, which provides a balanced assessment of model performance, especially on data with an unbalanced class distribution.

Mathematically, the calculation of the evaluation metric can be formulated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad 8), \text{Precission} = \frac{TP}{TP+FP} \quad 9), \text{Recall} = \frac{TP}{TP+FN} \quad 10), F1-\text{Score} = 2x\,\frac{\text{Precission x Recall}}{\text{Precission x Recall}} \quad 11)$$

In the context of this research, Confusion Matrix is used to assess the extent to which the SVM model trained with TF-IDF weighting is able to accurately classify the disease types in Lontar Usada Rare. This evaluation not only helps in assessing the overall accuracy of the model, but also provides a detailed description of the model's weaknesses, for example in classes with infrequent data. Thus, Confusion Matrix becomes an important component in ensuring the reliability of the model as well as its ability to generalize to new data.

## RESULT

The test results show that varying the complexity parameter C has a significant effect on the classification performance in each scenario of the training and test data ratio. In general, the value of C = 10 proved to consistently provide the highest accuracy compared to other values in all test ratios. For example, in the 80:20 ratio, the model accuracy only reached 74.12% at C = 0.5, increased to 82.35% at C = 1, and stabilized at around 87.06% at C = 10, 100, and 1000. A similar pattern is also found in the other ratios, where the value of C = 10 consistently performs the best, indicating that this parameter is able to provide an optimal balance between the decision boundary margin and the prediction error rate. Therefore, the value of C = 10 is set as the optimal parameter for the final model building.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

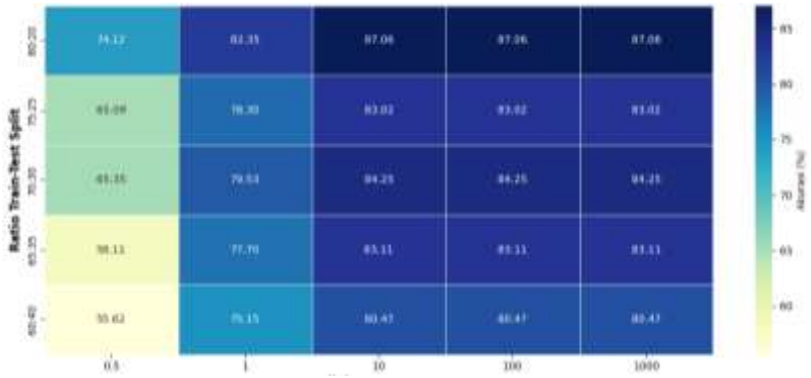

Fig. 2. Heatmap of SVM Model Accuracy Across Various Ratio and Parameter C Combinations

To comprehensively visualize the effect of parameter C and train-test split ratios on model performance, a heatmap analysis was conducted as shown in Fig. 2. The heatmap reveals that accuracy consistently increases from C = 0.5 to C = 10 across all split ratios, with the 80:20 scenario showing the most dramatic improvement from 74.12% to 87.06%. A plateau effect emerges at C ≥ 10, where accuracy stabilizes at 87.06% (80:20), 83.02% (75:25), 84.25% (70:30), 83.11% (65:35), and 80.47% (60:40), indicating that C = 10 represents optimal regularization strength. The darker blue cells concentrated in the upper rows (80:20, 75:25) contrast sharply with lighter cells in lower rows (60:40, 65:35), confirming that larger training proportions yield superior performance, with the optimal configuration at C = 10 and 80:20 ratio achieving the highest accuracy.

Table. 6. Model Testing Results

| Ratio Data | Kernel | C Optimal | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|---|
| 80:20 | Linear | 10 | 87.06% | 91.55% | 87.06% | 87.96% |
| 75:25 | Linear | 10 | 83.02% | 86.37% | 83.02% | 83.38% |
| 70:30 | Linear | 10 | 84.25% | 86.18% | 84.25% | 84.01% |
| 65:35 | Linear | 10 | 83.11% | 85.39% | 83.11% | 82.69% |
| 60:40 | Linear | 10 | 80.47% | 83.61% | 80.47% | 80.50% |

In more detail, the best results were obtained at a ratio of 80:20, with 87.06% accuracy, 91.55% precision, 87.06% recall, and 87.96% F1-score. The high precision value reflects the model's ability to minimize false positive predictions, while the high recall confirms the model's effectiveness in detecting truly relevant data. F1-score, which is balanced with precision and recall, shows that the model is not only accurate, but also consistent in handling variations in test data. Thus, it can be concluded that the combination of 80:20 ratio and parameter C = 10 is the most optimal configuration in this study.

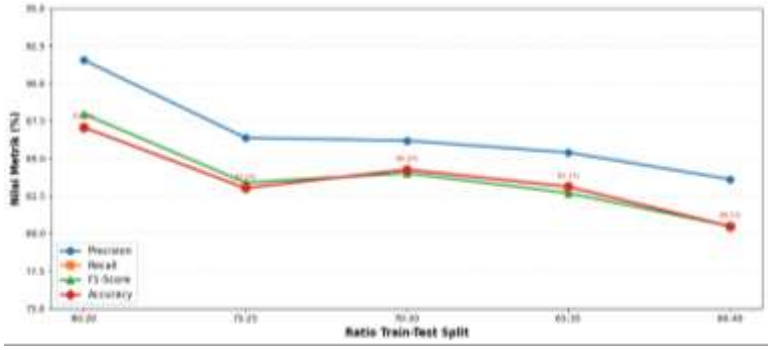

Fig. 3. Comparison of SVM Model Evaluation Metrics (C=10) Across Various Train-Test Split Ratios

A deeper examination of evaluation metrics across different train-test split ratios at C = 10 is presented in Fig. 3. The line graph shows that precision exhibits the steepest decline from 91.55% (80:20) to 83.61% (60:40), while recall and accuracy follow nearly identical trajectories from 87.06% to 80.47%. F1-score maintains a balanced decline from 87.96% to 80.50%, reflecting proportional degradation in both precision and recall. Notably, all four metrics converge at approximately 80-83% at the 60:40 ratio, indicating that reduced training data constrains the model's discriminative capacity. The consistent gap between precision and other metrics across all ratios demonstrates that the 80:20 configuration achieves superior performance with the highest precision (91.55%) while maintaining strong recall (87.06%), validating this ratio as optimal for the final model.

After selecting the optimal configuration, prediction testing was conducted using the linear SVM model with an 80:20 data split and parameter C = 10. This evaluation assesses the model’s ability to predict pediatric disease

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

types based on user-entered symptom text. Each symptom input is converted into numerical features using TF-IDF and then classified by the SVM model to generate the most relevant disease label along with corresponding herbal ingredients and treatment procedures derived from the Lontar Usada Rare text. This process demonstrates the practical implementation of the final classification model for interpreting traditional medical knowledge.

Fig. 2. Prediction Test 1 Result

In the first test, the user entered symptoms such as "the baby's eyes look blue on the whites, the hands and feet feel cold, the body relaxes and dims, and there is a split in the fontanel." Based on the results of text processing using the Term Frequency-Inverse Document Frequency (TF-IDF) method and classification with a linear kernel Support Vector Machine (SVM), the system successfully predicted Hemisphere disease correctly. The model then displays two traditional treatment recommendations taken from the Lontar Usada Rare manuscript, each containing natural ingredients such as lotus bulbs, garlic, and red sticky rice, along with the procedures for their use.

Fig. 3. Prediction Test 2 Result

In the second test, the user entered the symptoms "yellow-white bell in baby's eyes, uneven energy, and porous lips". Based on the processing results using TF-IDF and classification by linear kernel SVM, the system successfully predicted the Jampi disease. Interestingly, the model not only provided a single result, but also displayed several relevant traditional treatment recommendations from Lontar Usada Rare. Some of these included ingredients such as sembung root, toasted coconut, and garlic, complete with processing procedures. This shows that the model is able to recognize semantic relationships between symptoms and associate them with several variations of treatment recipes within the same disease class. Thus, the system not only performs precise classification, but also provides richer and contextualized outputs characteristic of Balinese traditional medical knowledge.

Fig. 4. Prediction Test 3 Result

In the third test, the entered symptoms of "slightly white eyes, weak energy, and high fever" were successfully predicted as Guaman disease. The model then displays the medicinal ingredients such as betel nut, pepper, and Chinese pijer, complete with how to process them. This prediction accuracy demonstrates the effectiveness of the combination of TF-IDF and linear kernel SVM in identifying symptom terms that are semantically similar to the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15508
e-ISSN : 2541-2019
p-ISSN : 2541-044X

training data, despite being syntactically different. This proves that the method used has good generalization ability to variations in traditional text data.

Based on three prediction test scenarios, the Support Vector Machine (SVM) model with linear kernel and optimal parameter C = 10 at an 80:20 ratio demonstrated excellent performance in recognizing pediatric disease symptom patterns from Lontar Usada Rare text. Despite variations in symptom sentence structure and vocabulary across tests, the system accurately classified disease types and generated multiple traditional treatment recommendations, including herbal ingredients and processing procedures sourced directly from the manuscript. This validates that the TF-IDF and linear kernel SVM combination achieves optimal classification performance with 87.06% accuracy, effectively capturing semantic relationships between symptoms and disease contexts. The integrated pipeline from preprocessing through TF-IDF weighting to SVM classification successfully transforms complex traditional text into structured outputs, demonstrating the practical potential of machine learning in digitizing and modernizing Balinese traditional medical knowledge for broader contemporary healthcare applications.

## DISCUSSIONS

The test results show that the Support Vector Machine (SVM) model with linear kernel trained using TF-IDF features achieves the best performance in the 80:20 data splitting scenario with complexity parameter C = 10. In this scenario, the model obtained an accuracy value of 87.06%, precision 91.55%, recall 87.06%, and F1-score 87.96%. Compared to the other four ratio scenarios, these results demonstrate consistent and significant improvement. The combination of 80:20 ratio and parameter C = 10 produces the optimal balance between generalization ability and prediction accuracy.

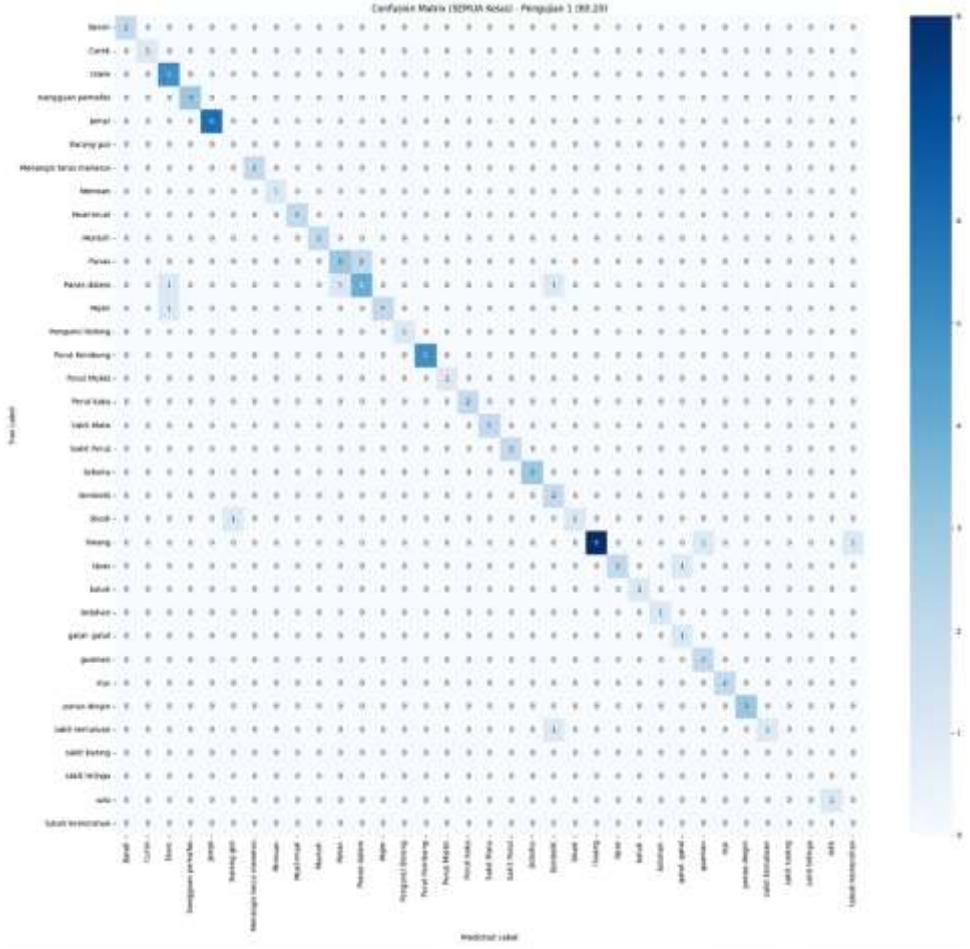

Fig. 5. Confusion Matrix Results

The Confusion Matrix visualization in Fig. 5 reveals that most disease classes were correctly predicted, with strong performance in Tiwang (8 correct predictions), Jampi (6), and Respiratory disorders (5). The dominant diagonal pattern indicates effective disease recognition. However, detailed error analysis reveals three primary misclassification patterns that warrant discussion. First, diseases with overlapping symptomatology exhibit cross-classification errors. For example, "Panas" (Heat), "Panas Dalam" (Internal Heat), and "Demam" (Fever) are

*name of corresponding author

frequently confused due to shared symptom descriptors in traditional text such as "badan panas" (hot body) and "suhu tinggi" (high temperature). This semantic similarity poses challenges for TF-IDF-based feature extraction, as these terms generate overlapping feature vectors that complicate class discrimination. Second, minority classes with fewer than 5 training samples demonstrate substantially lower recall rates. Examples include "Sakit Kuning" (Jaundice) and "Sakit Telinga" (Earache), each with only 2 samples. Insufficient training examples prevent the model from learning discriminative features, a limitation consistent with class imbalance challenges documented in medical text classification literature (Lu et al., 2022; Zhu et al., 2024). This underrepresentation directly impacts the model's ability to generalize to these rare disease categories. Third, linguistic factors contribute significantly to misclassification. The Lontar Usada Rare manuscript contains dialectal variations where identical symptoms appear with different spellings. For instance, "muntah," "metaah," and "mutah" all refer to vomiting. While normalization processes address many variations, rare dialectal forms remain unrecognized, leading to feature sparsity and subsequent classification errors.

To address the observed class imbalance problem, two promising techniques merit consideration in future research. First, Synthetic Minority Over-sampling Technique (SMOTE) could generate synthetic samples for minority classes by interpolating between existing instances in the feature space. For diseases like "Sakit Kuning" and "Sakit Telinga" with only 2 samples each, SMOTE would create additional training examples by identifying similar symptom patterns and generating new synthetic instances, enabling the model to learn more robust decision boundaries for underrepresented classes. Second, cost-sensitive learning through weighted SVM could assign higher misclassification costs to minority classes, forcing the model to prioritize learning patterns from rare diseases. This could be implemented using the class_weight='balanced' parameter or by manually defining class weights inversely proportional to class frequencies. For example, "Sakit Kuning" with 2 samples could be assigned a weight 13 times higher than "Diare" with 26 samples. These techniques have demonstrated effectiveness in medical text classification tasks with imbalanced data and represent practical directions for improving minority class performance.

To contextualize model performance, we compare our results with (Dwi Fasnuari et al., 2022), who applied K-Nearest Neighbor (KNN) for diabetes mellitus classification using 135 samples, achieving 93% accuracy, 100% precision, 60% recall, and 75% F1-score. While KNN demonstrated high precision, the substantially lower recall (60%) indicates failure to identify a significant portion of actual positive cases. In contrast, our SVM model achieves more balanced metrics with 91.55% precision and 87.06% recall, yielding a superior F1-score of 87.96%. This advantage stems from fundamental algorithmic differences. KNN relies on distance calculations and suffers from the curse of dimensionality in high-dimensional spaces. Our TF-IDF representation generates 400+ dimensional feature vectors corresponding to the Lontar Usada Rare vocabulary size, where distance-based metrics become less meaningful. Conversely, SVM is explicitly designed to handle high-dimensional data through kernel methods and margin maximization. Traditional medical texts with non-uniform vocabulary, spelling variations, and archaic linguistic patterns create sparse and irregular feature distributions where KNN's local similarity assumption breaks down, whereas SVM's margin-based approach proves more robust. The achieved performance (87.06% accuracy, 91.55% precision, 87.06% recall, 87.96% F1-score) validates that combining TF-IDF with SVM effectively digitizes traditional Balinese medical knowledge from historical manuscripts. The optimal 80:20 split ratio confirms the effectiveness of this approach in handling complex and unstructured traditional text data, establishing a replicable framework for cultural heritage digitization projects involving historical medical manuscripts from other traditions.

## CONCLUSION

This research successfully developed a pediatric disease classification model based on traditional Balinese medical text from Lontar Usada Rare using the Support Vector Machine (SVM) method combined with TF-IDF weighting. Testing on 422 samples across 35 disease classes produced the best performance at an 80:20 train–test ratio with parameter C = 10, achieving 87.06% accuracy, 91.55% precision, 87.06% recall, and an F1-score of 87.96%. These results indicate that the SVM–TF-IDF approach is effective in capturing textual patterns within traditional medical descriptions.

Error analysis shows that misclassifications mainly occur due to overlapping symptoms across disease classes, limited training data in minority categories, and challenges in normalizing Old Balinese terms with varying dialects and spellings. A comparison with the K-Nearest Neighbor (KNN) method further demonstrates SVM's superiority, as SVM produces more balanced precision–recall values and higher F1-score. These findings confirm that SVM handles high-dimensional sparse features more effectively than distance-based models for this type of cultural-linguistic dataset.

Overall, this study contributes to the digital preservation of Balinese traditional medical knowledge by demonstrating that machine learning can be applied to process complex textual narratives found in classical manuscripts. The combination of TF-IDF and linear SVM proves to be an accurate and computationally efficient approach for classifying pediatric-related symptoms within traditional medical texts, offering a practical foundation for broader digital heritage applications in the health domain.

*name of corresponding author

Future work will focus on expanding the dataset through the integration of additional Balinese medical manuscripts such as Usada Taru Pramana, and Usada Tenung using a standardized metadata framework to ensure consistency across sources. To address limitations in linguistic variation and enhance model capability, future research will explore fine-tuning deep learning models such as multilingual BERT or IndoBERT. These transformer-based approaches are expected to capture richer semantic and contextual nuances in traditional Balinese medical language, ultimately enabling a more robust and scalable classification system for digitizing and preserving indigenous medical heritage.

## REFERENCES

Abdalla, H. B. (2022). A Brief Survey On Big Data: Technologies, Terminologies And Data-Intensive Applications. Journal Of Big Data, 9(1). Https://Doi.Org/10.1186/S40537-022-00659-3

Adnyana, P. E. S. (2020). Lontar Usada Rare : Memahami Kearifan Lokal Tradisional Bali Dalam Mendiagnosa Gejala Penyakit Anak. 3(2), 163–173.

Adnyana, P. E. S. (2021). Empirisme Penggunaan Tumbuhan Pada Pengobatan Tradisional Bali: Lontar Taru Pramana Dalam Konstruksi Filsafat Ilmu. Sanjiwani: Jurnal Filsafat, 12(1), 64. Https://Doi.Org/10.25078/Sjf.V12i1.2059

Amaya-Tejera, N., Gamarra, M., Vélez, J. I., & Zurek, E. (2024). A Distance-Based Kernel For Classification Via Support Vector Machines. Frontiers In Artificial Intelligence, 7. Https://Doi.Org/10.3389/Frai.2024.1287875

Balipost.Com. (2024). Disbud Badung Lestarikan Lontar Dengan Digitalisasi. Https://Www.Balipost.Com/News/2024/02/08/387044/Disbud-Badung-Lestarikan-Lontar-Dengan...Html

Brin.Go.Id. (2025). Inovasi Digital Selamatkan Warisan Budaya Dan Bahasa Daerah. Https://Www.Brin.Go.Id/News/122398/Inovasi-Digital-Selamatkan-Warisan-Budaya-Dan-Bahasa-Daerah

Chen, W., Gong, Y., Xu, C., Hu, H., Yao, B., Wei, Z., Fan, Z., Hu, X., Zhou, B., Cheng, B., Jiang, D., & Duan, N. (2022). Contextual Fine-To-Coarse Distillation For Coarse-Grained Response Selection In Open-Domain Conversations. Proceedings Of The Annual Meeting Of The Association For Computational Linguistics, 1, 4865–4877. Https://Doi.Org/10.18653/V1/2022.Acl-Long.334

Dag, H. (2020). The Impact Of Text Preprocessing On The Prediction Of Review Ratings. May. Https://Doi.Org/10.3906/Elk-1907-46

Dewanti, T. R., Prathivi, R., & Susanto. (2025). Implementasi Metode Svm Untuk Klasifikasi Penyakit Stunting Bayi. 101–106.

Dwi Fasnuari, H. A., Yuana, H., & Chulkamdi, M. T. (2022). Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Penyakit Diabetes Melitus. Antivirus : Jurnal Ilmiah Teknik Informatika, 16(2), 133–142. Https://Doi.Org/10.35457/Antivirus.V16i2.2445

Fan, Q., Liu, S., Zhao, C., & Li, S. (2023). An Instance- And Label-Based Feature Selection Method In Classification Tasks. Information (Switzerland), 14(10), 1–14. Https://Doi.Org/10.3390/Info14100532

Fiebig, T., Litschko, C., Freiberger, F., Bethe, A., Berger, M., & Gerardy-Schahn, R. (2018). Cro Efficient Solid-Phase Synthesis Of Meningococcal Capsular Oligosaccharides Enables Simple And Fast Chemoenzymatic. 293, 953–962. Https://Doi.Org/10.1074/Jbc.Ra117.000488

Gastaldi, J. L., Terilla, J., Malagutti, L., Dusell, B., Vieira, T., & Cotterell, R. (2025). The Foundations Of Tokenization: Statistical And Computational Concerns. 1–18. Http://Arxiv.Org/Abs/2407.11606

Goodnewsfromindonesia.Id. (2025). Digitalisasi Lontar Bali Sebagai Upaya Menjaga Warisan Leluhur. Https://Www.Goodnewsfromindonesia.Id/2021/07/19/Digitalisasi-Lontar-Bali-Sebagai-Upaya-Menjaga-Warisan-Leluhur

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics For Multi-Class Classification: An Overview. 1–17. Http://Arxiv.Org/Abs/2008.05756

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An Overview On The Advancements Of Support Vector Machine Models In Healthcare Applications: A Review. Information (Switzerland), 15(4). Https://Doi.Org/10.3390/Info15040235

Hidayat, S., Napitupulu, H., & Gusriani, N. (2024). Penerapan Model Support Vector Machine Pada Kasus Klasifikasi Teks Berdasarkan Tujuan Sdgs Ke Tiga, Empat, Dan Enam. 6(2), 28–37.

Khan, N., Elizondo, D., Deka, L., & Molina-Cabello, M. A. (2024). Natural Language Processing Tools And Workflows For Improving Research Processes. Applied Sciences (Switzerland), 14(24), 1–26. Https://Doi.Org/10.3390/App142411731

Kim, H., You, S., Park, Y., Choi, J. Y., Ma, Y., Hong, K. T., Koh, K. N., Yun, S., Lee, K. H., & Shin, H. Y. (2021). Interplay Between Il6 And Crim1 In Thiopurine Intolerance Due To Hematological Toxicity In Leukemic Patients With Wild - Type Nudt15 And Tpmt. Scientific Reports, 1–13. Https://Doi.Org/10.1038/S41598-021-88963-5

Kunilovskaya, M. (2021). Text Preprocessing And Its Implications In A Digital Humanities Project. 85–93.

Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022). A Comparative Study On Deep Learning Models For Text Classification Of Unstructured Medical Notes With Various Levels Of Class Imbalance. Bmc Medical Research Methodology, 1–12. Https://Doi.Org/10.1186/S12874-022-01665-Y

Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., & Borodulin, A. (2025). Support Vector Machines In Polymer Science: A Review. Polymers, 17(4), 1–26. Https://Doi.Org/10.3390/Polym17040491

Navarro, C. F., & Perez, C. A. (2019). Color-Texture Pattern Classification Using Global-Local Feature Extraction, An Svm Classifier, With Bagging Ensemble Post-Processing. Applied Sciences (Switzerland), 9(15), 9–16. Https://Doi.Org/10.3390/App9153130

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Van Le, H., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence Of

*name of corresponding author

Data Splitting On Performance Of Machine Learning Models In Prediction Of Shear Strength Of Soil. Mathematical Problems In Engineering, 2021. Https://Doi.Org/10.1155/2021/4832864

Peraturan Menteri Kesehatan Nomor 37 Tahun 2017 Tentang Pelayanan Kesehatan Tradisional Integrasi, Pub. L. No. 1109/Menkes/Per/Ix/2007 (2017).

Pradana, A. W., & Hayaty, M. (2019). The Effect Of Stemming And Removal Of Stopwords On The Accuracy Of Sentiment Analysis On Indonesian-Language Texts. 4(3).

Putra Asana, I. M. D., & Della Tirta Yanti, N. P. (2023). Sistem Klasifikasi Pengajuan Kredit Dengan Metode Support Vector Machine ( Svm ). 06(02), 123–133.

Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect Of Dataset Size And Train/Test Split Ratios In Qsar/Qspr Multiclass Classification. Molecules, 26(4), 1–16. Https://Doi.Org/10.3390/Molecules26041111

Rahayu, S., & Yamasari, Y. (2024). Klasifikasi Penyakit Stroke Dengan Metode Support Vector Machine (Svm). Journal Of Informatics And Computer Science (Jinacs), 5(03), 440–446. Https://Doi.Org/10.26740/Jinacs.V5n03.P440-446

Ropikoh, I. A., Abdulhakim, R., Enri, U., & Sulistiyowati, N. (2021). Penerapan Algoritma Support Vector Machine ( Svm ) Untuk Klasifikasi Berita Hoax Covid-19. 5(1).

Saputra, N. A., Aeni, K., & Saraswati, N. M. (2024). Indonesian Hate Speech Text Classification Using Improved K-Nearest Neighbor With Tf-Idf- Icsρf. 11(1), 21–30. Https://Doi.Org/10.15294/Sji.V11i1.48085

Sarica, I. S., & Luo, J. (2021). Stopwords In Technical Language Processing. 1–13. Https://Doi.Org/10.1371/Journal.Pone.0254937

Sathyanarayanan, S., & Roopashri Tantri, B. (2024). Confusion Matrix-Based Performance Evaluation Metrics. African Journal Of Biomedical Research, 27(4), 4023–4031. Https://Doi.Org/10.53555/Ajbr.V27i4s.4345

Sheridan, P., Ahmed, Z., & Farooque, A. A. (2025). A Fisher's Exact Test Justification Of The Tf–Idf Term-Weighting Scheme. The American Statistician, 1–11. Https://Doi.Org/10.1080/00031305.2025.2539241

Sun, J., Li, L., Wang, P., Zhang, S., & Wu, J. (2017). And Expression Analysis Of The Leucine-Rich Repeat Receptor-Like Protein Kinase ( Lrr-Rlk ) Gene Family In Rosaceae Genomes. 1–15. Https://Doi.Org/10.1186/S12864-017-4155-Y

Uu Nomor 5 Tahun 2017 Tentang Pemajuan Kebudayaan, Pub. L. No. Uu Nomor 5 Tahun 2017 (2017).

Venkataraman, G. R., Pineda, A. L., Bear, O. J., Iv, W., Zehnder, A. M., Ayyar, S., Page, R. L., Bustamante, C. D., & Id, A. R. (2020). Plos One Fastag : Automatic Text Classification Of Unstructured Medical Narratives. 1–18. Https://Doi.Org/10.1371/Journal.Pone.0234647

Verdikha, N. A., & Yulianto, F. (2025). Klasifikasi Ulasan Aplikasi Sirekap 2024 Dengan Ekstraksi Fitur Word2vec Dan Metode Support Vector Machine ( Svm ). 9(2), 3013–3019.

Wabula, Y., Latief, A. D., & Zainuddin, H. (2023). Next Sentence Prediction : The Impact Of Preprocessing Techniques In Deep Learning. 2023 International Conference On Computer, Control, Informatics And Its Applications (Ic3ina), October, 274–278. Https://Doi.Org/10.1109/Ic3ina60834.2023.10285805

Weng, W., Wagholikar, K. B., Mccray, A. T., Szolovits, P., & Chueh, H. C. (2017). Medical Subdomain Classification Of Clinical Notes Using A Machine Learning-Based Natural Language Processing Approach. 1–13. Https://Doi.Org/10.1186/S12911-017-0556-8

Zhu, J., Pu, S., He, J., Su, D., Cai, W., Xu, X., & Liu, H. (2024). Processing Imbalanced Medical Data At The Data Level With Assisted-Reproduction Data As An Example.

*name of corresponding author