CHAPTER I
INTRODUCTION

1.1. Background

Language is the primary means by which humans communicate and convey ideas to others. In the context of social and cultural life, language also serves as a social group's identity [1]. In Indonesia, regional languages are an important part of local cultural identity that needs to be preserved [2]. However, a number of statistics show that the use of regional languages in Indonesia continues to decline.

The decline in regional language use is evident in the results of the 2020 Population Census, which showed that 74.77 percent of Indonesians aged five and over used regional languages to communicate within their families, down from 79.64 percent in the 2010 Census. A similar trend was seen in the community (72.78 percent in the 2020 Census). Furthermore, this decline is generational: the younger a generation, the less likely it is to use regional languages within the family [3].

This phenomenon indicates a shift in intergenerational communication patterns that could potentially threaten the sustainability of regional languages, including Javanese, the most widely spoken language in Indonesia. According to data from the Central Statistics Agency (2020, Long Form), Javanese is the most widely spoken language. Among the population aged 5 and over who speak regional languages, 42.12 percent speak Javanese within their families [3]. Despite its numerical dominance, the general trend of declining use of regional languages indicates that Javanese also faces similar challenges, particularly in the process of intergenerational inheritance. This situation is further complicated by the vast diversity of Javanese dialects, which require special attention in preservation and documentation efforts [4].

Linguistic mapping of Javanese language traditionally often refers to the three main groups proposed by Uhlenbeck [5]. He identified three main dialect groups, namely: ngapak (covering Banten, Tegal, Banyumas, Cirebon), tengahan (covering Pekalongan, Semarang, Yogyakarta, Madiun), and timuran (covering the Pantura of East Java, Surabaya, Malang, Jombang, Banyuwangi, and Tengger). This three-group classification is indeed useful in general, but risks ignoring the phonetic variations that exist in more specific sub-dialects, which are crucial to document in comprehensive digital preservation efforts.

Amid these challenges, digital media technology has emerged as a crucial tool in language preservation efforts in the modern era. According to Hinton et al., this approach enables meticulous documentation, systematic archiving, and digitization of language, expanding data accessibility. Furthermore, digital technology provides tools for capturing and preserving oral traditions, songs, or stories in high-quality formats. Similarly, various Artificial Intelligence (AI)-based tools are now also being utilized in efforts to preserve and analyze regional languages [6].

However, the application of computational methods for classifying language variations in Java still faces significant challenges. Fauzi et al. conducted a study to identify language variations in Java, including Betawi, Sundanese, Banyumasan, and Suroboyoan. The study used MFCC feature extraction and ANFIS classification, but reported a very low average accuracy of 32.5% [7]This low performance occurred even though the objects studied had quite contrasting language differences, indicating the need for more reliable methods.

This challenge is even more evident in the classification of internal Javanese dialects, where differences are very subtle. A recent comparative acoustic study in 2024 revealed that speakers of the Pekalongan dialect have a much higher and more strident tone than the heavier Pati dialect, and have a distinctive 'cengkok' at the end of sentences [8]. In addition to the variations in tone, the differences in tempo are also very significant; East Javanese dialects (such as Surabaya) are proven to have a much faster and more dynamic speech rhythm than Central Javanese dialects, which tend to be slower [9].

Based on the existing findings, it can be concluded that the characteristics of Javanese dialects are stored in two main dimensions: the frequency dimension, which reflects the pitch and the temporal dimension, which relates to the speed or slowness of the tempo. This indicates that modeling strategies must be able to accommodate both spectral details and durational dynamics in a balanced manner for optimal dialect detection.

One of the most reliable methods for capturing detailed audio features is the Convolutional Neural Network (CNN). CNNs are widely used for image classification but can also be applied to speech recognition. In practice, CNNs are very effective in extracting features from spectrograms, particularly Mel-spectrograms, which are capable of providing descriptive sound representations that closely resemble human

auditory perception. This representation makes Mel-spectrograms particularly suitable for recognizing dialects with frequency differences, such as differences in pitch between speakers of the Javanese dialect. Furthermore, CNNs have advantages in handling noisy or low-resolution audio data [10]. However, because CNNs treat the input as a two-dimensional representation like a static image, this method has limitations in explicitly modeling long-term temporal dependencies [11], even though this aspect plays an important role in forming dialect intonation patterns.

To address these shortcomings in the temporal aspect, this study applies a Temporal Convolutional Network (TCN) as a sequential data modeler. TCN has a specific advantage in detecting temporal features that are often missed by standard convolutional architectures. The TCN architecture is designed with a core block capable of recognizing long-term dependencies in speech signals, enabling it to effectively capture information from a wide time span. This capability is vital in audio analysis because speech signals contain dynamic elements such as tone, stress, and speed that reveal the speaker's characteristics [12]. In the context of Javanese dialect, the application of TCN aims to capture prosodic variations such as 'cengkok' patterns or rapid changes in speech tempo so that the system can understand the dialect context as a whole, complementing the static acoustic features extracted by CNN.

Based on the analysis of the characteristics of the above methods, this study proposes the implementation of a CNN-TCN hybrid architecture as an integrative approach. In this design, CNN functions as a front-end component that extracts deep spatial acoustic features from Mel-spectrograms, then the extraction results are forwarded to TCN as a back-end to model long-term temporal dependencies in speech signals. This approach is in line with the research of Chen et al. who showed that the use of CNN as a feature extractor followed by TCN can improve prediction accuracy on complex audio data [13].

This finding was also reinforced by Jo and Kwak in a study of speech signal analysis, which confirmed that the integration of spatial and temporal feature modeling was more effective in capturing dynamic prosodic elements, such as pitch and speech rate, compared to using a single model [12]. Through this hybrid approach, the system is expected to produce a more robust dialect classification, not only computationally superior but also more sensitive to the linguistic nuances unique to Javanese.

1.2. Formulation of the problem

The formulation of the research problem is formulated in the form of research questions as follows:

1. How to design and implement a hybrid CNN-TCN deep learning architecture to classify Javanese dialects based on Log-Mel Spectrogram input?
2. How high is the accuracy and performance of the proposed CNN-TCN model in identifying and distinguishing various Javanese dialects?

1.3. Research purposes

Based on the formulation of the problem that has been determined, the objectives of this research are:

1. Designing and implementing a deep learning model with a hybrid CNN-TCN architecture that is optimal for classifying Javanese dialects based on Log-Mel Spectrogram input.
2. Analyze and measure the performance of the proposed CNN-TCN model in identifying and distinguishing various dialects of Javanese language based on evaluation metrics such as accuracy, precision, recall, and F1-score.

1.4. Benefits of research

This research is expected to provide a series of significant benefits, both from a scientific perspective and practical application, which include:

1. Contributes to the field of computational dialectology by presenting an objective and measurable method for dialect analysis and classification, which can support modern language mapping studies.
2. To be a reference and case study for the development of computer science, especially in the application of deep learning for voice classification tasks in the domain of regional languages in Indonesia.
3. Produce a classification model that functions as a form of digital documentation of the richness of Javanese dialects, so that it can be used as a tool by cultural institutions in language preservation efforts.
4. Provides a fundamental foundation for the development of future, more dialect-aware voice-based technologies, such as educational apps, translation services, or more accurate speech-to-text systems for local speakers.

1.5. Scope of problem

To keep this research focused and directed towards the stated objectives, the scope of this research is limited by the following matters:

1. This study uses audio data sourced from the Javanese corpus (various dialects) available at The Language Archive (TLA) owned by the Max Planck Institute for Psycholinguistics as the main data reference, and is supplemented with additional data from public digital media platforms (YouTube).
2. This research limits the scope of classification to nine specific Javanese dialects: Arekan, Pantura (Banten), Banyumas, Cirebon, Mataraman, Osing, Solo-Yogya Tegal, and Tengger.
3. This research focuses on the task of dialect classification (Dialect Identification/DID). It does not include Automatic Speech Recognition (ASR) or speech transcription (converting speech to text).