DSTER: A Dual-Stream Transformer-based Emotion Recognition Model through Keystrokes Dynamics

Frank (Sicong) Chen, Shruti Rao, Brijesh Tiwari, Vir V. Phoha
Department of Electrical Engineering and Computer Science, Syracuse University
Syracuse, New York, 13210, USA
{schen154,srao18, brtiwari, vvphoha}@syr.edu

Abstract

Emotion Recognition is a critical research area for enhancing human-computer interaction. Keystroke dynamics, a behavioral biometric capturing typing patterns, offers a non-intrusive, user-friendly method for recognizing emotions. We propose a Dual-Stream Transformer-based Emotion Recognition (DSTER) model, which leverages keystroke dynamics to determine emotional states. The DSTER model features a dual-stream architecture that separately extracts temporal-over-channel and channel-over-temporal information. Each stream employs multi-head self-attention mechanisms, Long-Short Term Memory (LSTM), and Convolutional Neural Network (CNN) layers, along with dense vector embeddings of keycode data, to improve the extraction of temporal and contextual information from typing sequences. To the best of our knowledge, the DSTER model is the first to integrate transformer architecture with keystroke dynamics for emotion recognition. Our experiments on a widely-used fixed-text dataset demonstrate that the DSTER model significantly outperforms the three most recent baseline models, achieving average F1 scores up to 0.989 and an average accuracy increase of up to 66.04%. Unlike the significant performance variations reported in baseline models, the DSTER model maintains consistent and robust performance across all five tested emotional states. Further analysis shows that the model performs better with longer window lengths and greater overlaps.

1. Introduction

Advancements in digital devices necessitate improved human-computer interaction (HCI) to enhance user engagement and satisfaction. One of the most effective methods to enhance HCI is enabling the system to understand and adapt to the user's emotional context. This adaptation allows the system to respond appropriately, improving user satisfaction. Furthermore, emotion recognition is crucial in applications such as health monitoring [27] and healing [4] since understanding users' emotional states can greatly influence the effectiveness of interventions, such as the intensity and when it should intervene to help the patient recover from an unhealthy condition.

Emotion recognition (ER) has been a hot research area in the field of HCI for many decades [5, 20], and many studies have investigated ER through various biometrics, including physiological biometrics, such as facial expression [2, 8, 10] and Electrocardiogram (ECG) signals [7, 22], as well as behavioral biometrics, such as speech [6, 12] and keystroke dynamics [16, 26, 31]. Keystroke dynamics (KD), which refer to the manner and rhythm of typing, are less intrusive and require no additional permissions or invasive hardware, enhancing their suitability for ER from both usability and privacy perspectives. Moreover, with desktops and laptops being integral to daily life and continuously interacted with, implementing ER through KD on these devices allows for the development of personalized, context-aware applications that enhance user experience in a naturalistic setting [15].

In this paper, we propose a Dual-Stream Transformer-based Emotion Recognition (DSTER) model that utilizes keystroke dynamics for emotion recognition. The DSTER model leverages a dual-stream architecture that includes: (1) a temporal stream, focusing on extracting hidden sequential information and interrelationships between consecutive keys, and (2) a channel stream, concentrating on extracting inter-feature relationships within key and timing features. The input to these streams incorporates Gaussian Range Encoding (GRE) to assist the multi-head self-attention mechanism in understanding positional and contextual information. Additionally, Keystroke Dynamics Emotion Recognition (KDER) Multi-scale Long-Short Term Memory (Multi-LSTM) blocks and KDER Multi-scale Convolutional Neural Network (Multi-CNN) blocks in the temporal and channel streams are designed to further extract information from temporal and channel aspects. The

embeddings extracted from these streams are concatenated and used to make the emotion recognition decision.

We employed a widely-used KD dataset, EmoSurv dataset [17], in the experiment to evaluate the DSTER model and compared it against the three most recent studies in this field. The DSTER model demonstrated superior performance, achieving an accuracy of 99.44% and an F1 score of 0.989 across all five tested emotions. Compared to baseline studies that performed experiments on the same dataset, our model demonstrated significant improvements in accuracy, precision, recall, and F1 score. Notably, our model increased the average accuracy by a range from 22.96% to 66.04% over baseline models. Moreover, the DSTER model achieves consistent and robust performance across all emotional states, outperforming baseline models that reported significant performance variations across different emotions. Further analysis on different sample lengths and proportions of overlap suggests that our model performs optimally with longer samples and being trained on samples with greater overlap. The best performance is achieved at a sample length of 40 keystrokes and 90% overlap between samples. These results confirm that the DSTER model excels in multi-emotion recognition and is not user-specific but universally applicable, significantly enhancing its generalizability and suitability for real-world applications.

The main contributions of this paper are:

1. We propose a Dual-Stream Transformer-based Emotion Recognition (DSTER) model for emotion recognition through KD. To the best of our knowledge, we are the first to integrate transformer architecture into emotion recognition using keystroke data.

2. The dual-stream architecture consists of: (1) a temporal stream, which focuses on extracting sequential information between consecutive key vectors, and (2) a channel stream, which concentrates on unveiling hidden relationships between key and timing features.

3. Each stream incorporates Gaussian Range Encoding (GRE) to capture the uncertainty and influence of each position over its neighbors, enhancing the transformer's capability to process sequential information.

4. The DSTER model significantly outperforms recent studies and maintains consistent and robust performance across various emotional states, thus avoiding the considerable performance variations observed in baseline models.

5. The DSTER model is universal, applicable to any user rather than being user-specific, which enhances its practical applicability and deployment potential.

6. We investigate the impact of sample length and overlap proportion on model performance, identifying optimal settings for emotion recognition.

The remainder of this paper is organized as follows: Section 2 summarizes previous studies in emotion recognition, especially through keystroke dynamics, identifies their limitations, and briefly outlines how we addressed them. Section 3 details the feature extraction and architecture of our DSTER model. Section 4 presents experimental results, including comparisons with baseline models and analysis of how different sample configurations influence model performance. Finally, Section 5 concludes our work and discusses future research directions.

2. Prior work

2.1. Prior studies on emotion recognition through various biometrics

Emotion recognition (ER) has been extensively investigated over the past few decades. Researchers have explored emotional state recognition using physiological biometrics. For instance, Jain *et al*. [8] proposed a hybrid deep neural network (DNN) for facial emotion recognition, achieving an accuracy of 94.91%. Sarkar *et al*. [22] employed a series of convolutional neural network (CNN) blocks to learn representations from ECG signals, training a classification model that detected stress with an accuracy of 96.9%. Lee *et al*. [14] utilized a CNN model to extract features from photoplethysmogram (PPG) signals, classifying valence and arousal with accuracies of 75.3% and 76.2%, respectively. However, ER using these physiological biometrics typically requires additional devices to capture the signals, which must then be transferred to another device, such as a mobile phone or computer, for processing and analysis. It not only consumes time but also raises privacy concerns.

In contrast, behavioral biometrics can be captured in a more non-invasive manner and are often collected directly by the device with which users interact. For example, Yang *et al*. [30] developed an attention-based LSTM system that integrated signals from smartphone and wristband sensors to determine users' emotional states, achieving an average accuracy of 89.2% for binary positive and negative emotion classification. Xu *et al*. [29] enhanced speech ER accuracy by 6% over the previous state-of-the-art models using an attention-based CNN model. With advancements in large language models, text-based ER analysis has also gained prominence. Kumar *et al*. [13] proposed a text emotion recognition system that employed the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to extract embedding vectors, constructing a dual-channel neural network for emotion recognition and achieving an accuracy of 79.17%. Nonetheless, due to privacy

concerns, individuals are often hesitant to allow applications to access their speech or text data, making it hard to recognize their emotional state through these biometrics.

2.2. Prior studies on emotion recognition through keystroke dynamics and their limitations

Keystroke dynamics (KD) offer a less intrusive biometric for ER, without concerns over the content of the typing. For example, Kołakowska [11] investigated suitable classification models for ER using KD on physical keyboards, finding that accuracies varied significantly among emotional states and models, ranging from 47.36% to 81.25%. Qi *et al*. [21] analyzed positive and negative impulses from key presses and releases on a piezoelectric touch panel and extracted several time-related features. They constructed a random forest model that recognized four emotional states (happiness, fear, disgust, and sadness) with an accuracy of 78.31%. However, the small dataset size of only nine individuals may not fully substantiate their model's generalizability. Velichko and Izotov [28] evaluated an emotion recognition model based on the LogNNet neural network using the EmoSurv dataset [17], but achieved only a 33.4% accuracy across five emotional states (happy, sad, angry, calm, and neutral) with an F1 score of 0.299. Although their model required minimal computing resources, its performance was relatively low due to the simplicity of their model structure. Marrone and Sansone [19] proposed CNN and Multi-Instance Learning Support Vector Machine (MIL-SVM) based models, using a fixed time window of 15 seconds for data capture and emotion recognition. Their evaluations on the EmoSurv dataset showed that the MIL-SVM model performed better, achieving an average accuracy of 76%. Similarly, Maalej *et al*. [18] compared four traditional machine learning algorithms — J48, Random Forest, Random Committee, and KNN — on the same dataset, with accuracies peaking at 76.4818% and 76.5435% for fixed-text and free-text subset, respectively.

Nevertheless, existing models have not effectively capitalized on the sequential information inherent in a series of key presses and releases. Many of them extracted feature vectors from only two or three consecutive keystrokes, which may not be sufficient to accurately reveal emotional states. It also overlooks the potential interrelationships between consecutive actions. Moreover, they primarily rely on manually extracted features and simple machine learning models, which may hinder their ability to uncover deeper insights from the data necessary for enhanced performance. Furthermore, the accuracy of these models remains modest, peaking at approximately 76% for recognizing five emotional states and 78% for four in fixed-text typing scenarios. This modest performance underscores a significant opportunity for improvement.

2.3. How our model addresses the limitations

To overcome these limitations, we introduce the Dual-Stream Transformer-based Emotion Recognition (DSTER) model. This model processes a sequence of feature vectors as input, providing more comprehensive information than a single vector would. It utilizes a dual-stream architecture to separately extract temporal and contextual information. The temporal stream employs transformer architecture and Multi-LSTM blocks to extract sequential information, while the channel stream utilizes transformer architecture and Multi-CNN blocks to focus on extracting information within feature vectors. Previous studies [23, 25] have shown that the transformer works well in extracting information from KD features, and we incorporate Multi-LSTM and Multi-CNN to further extract information from each stream. Additionally, we integrate Gaussian range encoding (GRE) prior to processing by the dual-stream architecture, utilizing multiple Gaussian distributions to encode positions. Compared to traditional positional encoding, this approach better captures the uncertainty and influence of each position relative to its neighbors, thereby providing a more nuanced understanding of positional relationships. This enhancement is crucial for managing long sequences and improving the transformer's overall performance.

3. Dual-Stream Transformer-based Emotion Recognition Model

In this section, we provide a detailed description of the Dual-Stream Transformer-based Emotion Recognition (DSTER) model. We first explore the process of feature extraction from raw keystroke data, focusing on timing features and keycode embeddings. Then, we introduce the architecture of the DSTER model, presenting an in-depth analysis of each principal component.

3.1. Feature extraction for timing features and feature embedding for keycode

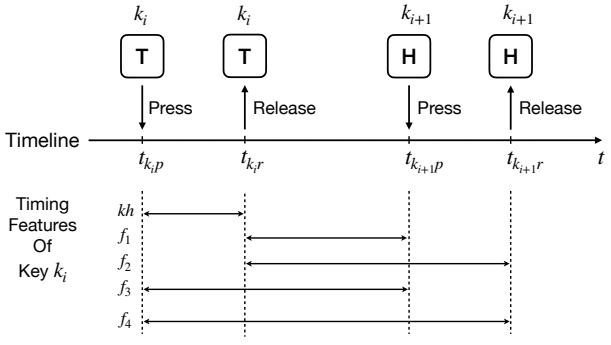

Figure 1. Illustrations of Timing Feature Extraction.

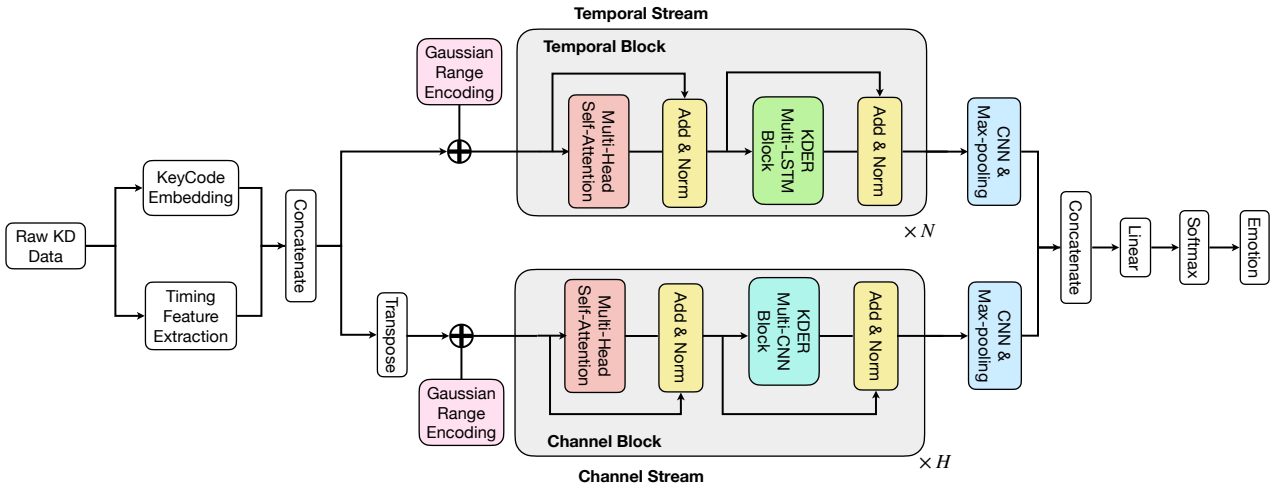

Figure 2. Architecture of Dual-Stream Transformer-based Emotion Recognition (DSTER) Model. (KDER: Keystroke dynamics emotion recognition, CNN: convolutional neural network, LSTM: long-short term memory.)

The raw typing data collected records the events when keys are pressed or released, including the keycode and timestamps. We extract five timing features from the raw data, as illustrated in Figure 1. For each pair of consecutive key presses, k_i and k_{i+1}, we utilize their press times (t_{k_ip} and $t_{k_{i+1}p}$) and release times (t_{k_ir} and $t_{k_{i+1}r}$) to compute the following timing features for key k_i: (1) Key hold time: $kh = t_{k_ir} - t_{k_ip}$; (2) Flight1 time: $f_1 = t_{k_{i+1}p} - t_{k_ir}$; (3) Flight2 time: $f_2 = t_{k_{i+1}r} - t_{k_ir}$; (4) Flight3 time: $f_4 = t_{k_{i+1}p} - t_{k_ip}$; (5) Flight4 time: $f_4 = t_{k_{i+1}r} - t_{k_ip}$.

Additionally, keycodes, which represent the keys pressed or released, are categorical variables. Unlike previous studies that directly used keycode values [24, 25] or one-hot encoded vector [3, 9], we integrate an embedding layer to encode each keycode to a dense vector space. This approach is not only more space-efficient compared to generating a sparse vector for each keycode through one-hot encoding, but it also allows the embedding layer to learn and capture potential informational and relational dynamics between consecutive keys.

The extracted timing features and embedded keycode vector are then concatenated to form the feature vector for key k_i as: [embedded keycode vector, kh, f_1, f_2, f_3, f_4]. The model then processes S consecutive feature vectors to recognize emotional states.

3.2. Architecture of Dual-Stream Transformer-based Emotion Recognition (DSTER) Model

The Dual-Stream Transformer-based Emotion Recognition (DSTER) model leverages a dual-stream structure underpinned by the multi-head self-attention mechanism. The input, comprising S consecutive key feature vectors as described in the previous section, is initially added with Gaussian range encoding. Subsequently, two streams of Transformer-based blocks process the input from different perspectives: the temporal stream, which extracts information and relationships between feature vectors, and the channel stream, which focuses on inter-feature relationships within each vector. Data processed by both streams are further analyzed through CNN layers with max-pooling to derive embeddings, which are concatenated and then processed through linear layers with a softmax activation function to recognize multiple emotional states. The overall model architecture is depicted in Figure 2. The rest of this section details each primary component of the model.

Gaussian Range Encoding (GRE): The sequence of S key feature vectors inherently contains valuable inter-sample relationships crucial for emotion recognition, as independent typing events do not signify emotions unless considered collectively. Traditional multi-head self-attention mechanisms treat input data as independent points and do not inherently capture sequential data relationships. To address this, we implement GRE [1] to enhance the transformer's ability to apprehend sequential information. GRE employs multiple Gaussian distributions with varied parameters to encode each position, enabling a position to belong to multiple ranges. This encoding approach allows the model to learn the extent of influence each range has on a given position. Compared to traditional positional encoding, it captures the uncertainty and influence of each position over its neighbors, thereby providing a more nuanced understanding of positional relationships and contextual information from sequential data. It thus enhances the transformer's capability to manage long sequences. GRE is applied to inputs for both the temporal and channel streams, as shown in Figure 2.

Temporal Stream: This stream consists of N identical temporal blocks, each comprising a residual-connected

multi-head self-attention layer, followed by layer normalization, and a residual-connected KDER (Keystroke Dynamics Emotion Recognition) Multi-LSTM (Multiple Long-short Term Memory) block, followed by another layer normalization. The multi-head self-attention mechanism integrates information across the entire input sequence, generating a hidden representation. This mechanism enables the model to simultaneously attend to different parts of the sequence and weigh the influence of all points, making it adept at capturing patterns that span various time steps. Following this, the KDER Multi-LSTM block employs a series of LSTM layers with varying hidden sizes to extract time-related features from this representation at different temporal scales. This multi-scale approach enables the model to capture both short-term dependencies and long-term trends, effectively extracting temporal dynamics crucial for accurate emotion recognition. Each LSTM layer is followed by ReLU activation, batch normalization, and a dropout layer. A fully connected layer then standardizes the output size to match the input for residual integration. Information extracted from various time scales is amalgamated by averaging the outputs from all LSTM layers as the final output of the KDER Multi-LSTM block.

Suppose the input of the KDER Multi-LSTM block is $X \in \mathbb{R}^{S \times d}$, and there are M LSTM layers with a set of hidden dimensions $M = \{m_1, m_2, ..., m_M\}$. The operations within a KDER Multi-LSTM block are mathematically described as follows:

$$T_{m_i} = \text{ReLU}(\text{Dropout}(\text{BN}(\text{ReLU}(\text{LSTM}(W_{m_i}, X))))) \tag{1}$$

$$O_{\text{LSTM}} = \frac{1}{M} \sum_{i=1}^{M} \text{FC}_{m_i}(T_{m_i}) \tag{2}$$

where T_{m_i} is the output of each LSTM layer and O_{LSTM} is the final output of the KDER Multi-LSTM block.

Channel Stream: The channel stream is designed to extract inter-feature information between encoded keycode features and timing features. The input is thus transposed, with rows representing features and columns representing time sequences. This stream contains H identical channel blocks, each comprising a residual-connected multi-head self-attention layer, followed by layer normalization, and a residual-connected KDER Multi-CNN (Multiple Convolutional Neural Network) block, also followed by layer normalization. The KDER Multi-CNN block employs multiple CNN layers with varying kernel sizes to capture channel-wise information at different scales. It enables the model to adaptively focus on varying extents of inter-feature relationships, from localized to more distributed patterns, providing a deeper analysis of how different features influence each other over time. Each CNN layer is followed by ReLU activation, batch normalization, and a dropout layer. The output from each CNN layer is summed and averaged as the block's output.

Suppose the input of the KDER Multi-CNN block is $X \in \mathbb{R}^{d \times S}$ and K is a set of kernel sizes of CNN layers $K = \{k_1, k_2, ..., k_K\}$. The output of the KDER Multi-CNN block O_{CNN} can be mathematically calculated as:

$$O_{\text{CNN}} = \frac{1}{K} \sum_{i=1}^{K} \text{ReLU}(\text{Dropout}(\text{BN}(\text{ReLU}(\text{CNN}(W_{k_i}, X))))) \tag{3}$$

Final Emotion Recognition: Following the processing through temporal stream and channel stream, a series of CNN and max-pooling layers are employed to extract embeddings from data processed by each stream. These embeddings are then concatenated, and a fully connected layer with a softmax activation function is utilized for the recognition of multiple emotional states.

4. Experiments and results

We used the widely recognized EmoSurv keystroke dataset [17] to evaluate our DSTER model and compare its performance against other studies. In this section, we first give a concise description of the EmoSurv dataset and our data preprocessing process. We then detail the training and evaluation procedures for our model. Subsequently, we present the experimental results obtained and compare these with those of baseline models. Additionally, we explore the DSTER model's performance under various parameter configurations, providing insights into how these configurations influence the model's effectiveness.

4.1. Deployment of DSTER model

4.2. Dataset Description and Pre-processing

Maalej *et al*. developed a dynamic web application, EmoSurv [17], specifically designed to collect keystroke dynamics data. During the data collection phase, participants initially provided typing data while in a neutral state. Subsequently, they were exposed to emotion-eliciting videos intended to induce a specific emotional state. After the emotion was induced, participants were asked to type again, with this subsequent data being annotated according to the induced emotional state.

The EmoSurv dataset encompasses five emotional states: happy, sad, angry, calm, and neutral, and includes both free-text and fixed-text typing data. For our experiments, we exclusively utilized the fixed-text typing data to recognize emotions, as this subset of data is commonly employed in emotion recognition studies [18, 19, 28]. The initial dataset included 83 participants; however, not all participants contributed data across all five emotional states. Following a

Study	Model	# emotions	Accuracy	Precision	Recall	F1-score
Marrone *et al*. [19]	MIL-SVM	5	76%	0.80	0.69	0.74
Velichko *et al*. [28]	LogNNet	5	33.4%	0.353	0.261	0.299
Maalej *et al*. [18]	Random Forest	4	76.4818%	-	-	-
Our work	DSTER	5	99.44%	0.986	0.992	0.989

Table 1. Comparison of average performance between our DSTER with baseline models. Performance from other studies are directly taken from corresponding studies.

rigorous data cleaning process to remove incomplete and noisy data, the remaining dataset comprised various numbers of participants and key feature vectors for each emotional state, detailed in Table 2.

Emotional States	Happy	Sad	Angry	Calm	Neutral
No. of Participants	32	29	23	27	82
No. of Feature vectors	4355	4508	4035	4668	25358

Table 2. Number of participants and number of key feature vectors used in our experiment.

4.2.1 Generation of training, validation, and testing datasets

To mitigate the potential overfitting due to single dataset evaluation, we partitioned participants across the training, validation, and testing datasets in an 8:1:1 ratio for each emotional state. This strategy ensures that no user's data overlaps between these sets, providing a robust validation through genuinely independent datasets for each phase of model evaluation. After that, Z-score normalization was applied separately to all samples based on the statistics from the training data. Sample generation leveraged the sliding window method with fixed-length sequences to select S consecutive key feature vectors and generate samples for each participant. We experimented with sliding window lengths S of [10, 20, 30, 40, 50, 60] and proportions of overlap in [90%, 80%, 70%, 60%, 50%]. It is worth noting that longer window lengths and smaller proportions of overlap yield fewer samples.

The hyperparameters for our model were set as follows: batch size of 16, learning rate of 0.001, loss function of cross-entropy, and the Adam optimizer. We incorporated an early stopping mechanism activated after 20 epochs without improvement. The model architecture included 10 temporal blocks ($N = 10$) and 10 channel blocks ($H = 10$), and the size of embedding layer for keycode embedding is set to 8.

4.2.2 Baseline models

To establish benchmarks, we compared our model against three studies within three years that also utilized the EmoSurv dataset with fixed-text data for emotion recognition: Marrone *et al*. [19], Velichko *et al*. [28], and Maalej *et al*. [18], as discussed in Section 2.2. We extracted performance metrics directly from these studies for comparison, which are summarized in Table 1 and 3.

4.3. Performance evaluation and insights

In this section, we present a comparative analysis of the best performance achieved by our DSTER model against baseline models. Additionally, we explore the performance variations of our DSTER model across different sample lengths and overlaps, analyzing how these factors influence recognition accuracy across various emotional states.

4.3.1 Performance evaluation and comparison with baseline models

Table 1 presents the average performance across all emotions for both our DSTER model and the baseline models. The baseline models' performances are as reported in their respective studies, while the performance of our DSTER model is the best achieved, utilizing a sliding window length of 40 keys and an overlap of 90% to generate samples. Notably, the DSTER model surpasses all baseline models, achieving an accuracy of 99.44% and an F1 score of 0.989. In contrast, the best-performing baseline model [19] achieved only 76% accuracy and an F1 score of 0.74. Thus, our DSTER model demonstrates a significant improvement, with a 23.44% higher accuracy than the baseline.

Study	Happy	Sad	Angry	Calm	Neutral
Marrone *et al*. [19]	0.78	0.71	0.58	0.60	0.94
Velichko *et al*. [28]	0.275	0.395	0.346	0.333	0.299
Maalej *et al*. [18]	0.796	0.815	0.717	0.723	-
Our work	0.981	0.997	0.989	0.979	0.999

Table 3. Comparison of average F1 score between our DSTER and baseline models for each emotion. Performance values from other studies are directly taken from corresponding studies.

Furthermore, Table 3 compares the performance of DSTER and baseline models across five emotional states, evaluated by F1 scores. All baseline models exhibited substantial variations in performance across different emotions. For instance, Velichko *et al*. [28] demonstrated poor performance on 'Happy,' achieving only an F1 score of 0.275, while Marrone *et al*. [19] underperformed on 'Angry' and 'Calm,' with F1 scores of only 0.58 and 0.6, respectively. The F1 score differences between the underperforming and

well-performing emotions for baseline models were as large as 0.2. In contrast, our DSTER model consistently achieved the highest performance across all five emotions, with average F1 scores nearing 1. The maximum F1 score difference within our model was only 0.02, demonstrating robust performance across various emotional states.

Several factors may contribute to the superior performance of the DSTER model:

(1) **Feature Extraction:** The three baseline models rely on manually extracted timing features, such as dwell times, flight times, and their statistical analyses. These simple features, which can be linearly combined, correlated, or inferred from sequential analysis, may potentially reduce their effectiveness due to their simplicity. Conversely, our model involves extracting only five basic timing features and embedding the keycode into a dense vector. This embedding process likely provides richer contextual information between keystrokes than merely using keycode values. Additionally, these timing features are then processed by the dual-stream architecture of the DSTER model, further enhancing the extraction and utilization of relevant but hidden information for emotion recognition.

(2) **Model Architecture:** The baseline models employ either traditional machine learning models like MIL-SVM and Random Forest, or simple neural network architectures like LogNNet, which are generally not optimized for handling sequential data. In contrast, our DSTER model utilizes multi-head self-attention mechanisms designed to unearth hidden relationships across entire sequences. Additionally, the GRE and the dual-stream architecture, which focuses on extracting temporal-over-channel and channel-over-temporal information, enhances the efficiency and accuracy of feature extraction.

(3) **Data Sampling Strategy:** The substantial overlap of 90% in our sliding window approach generates a larger number of samples for both training and testing. This high degree of sample redundancy provides the model with ample opportunity to learn from slightly varied versions of similar data, aiding in stabilizing and refining the learning process while enhancing the model's generalizability to new data. The results also demonstrates that this approach does not lead to overfitting but rather increases the model's generalizability, as evidenced by its robust performance on the unseen user's data in the testing set.

4.3.2 Performance evaluation and comparison with different data settings and across different emotional states

We further evaluated the DSTER model under varying window lengths and proportions of overlap between samples to investigate how different input configurations influence the model's performance. The average F1 scores across all emotions are depicted in Figure 3. Analysis of the figure reveals the following key findings:

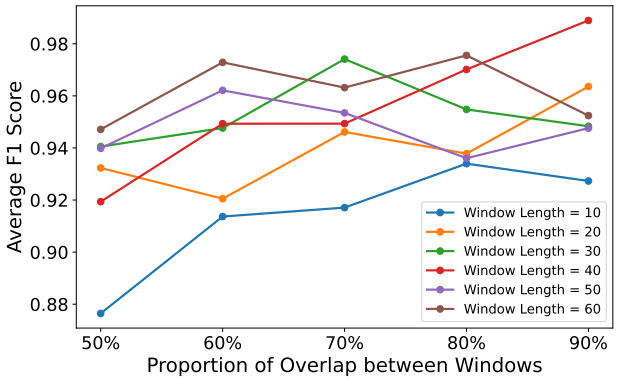

Figure 3. Average F1 scores of DSTER model across all emotions with different window lengths and proportions of overlaps.

(1) **Longer Window Length Yields Better Performance:** The model's overall performance improves with increased window length. Specifically, when the window length is set to 60, the performance metrics with different proportions of overlap are consistently above 0.94. In contrast, performance falls below 0.94 when the window length is reduced to 10. This improvement at longer window lengths can be attributed to the model's architecture, particularly via the multi-head self-attention mechanism, to capture long-range sequential information more effectively when more consecutive keys are involved. Furthermore, 10 or 20 key presses only correspond to approximately 2 - 5 words including spaces, and it is commonly understood that inferring emotions from just a few words can be challenging. Conversely, with 50 to 60 key presses, users may be able to type one or two complete sentences, which should be more conducive for accurate emotion recognition.

(2) **Being Trained on Samples with Larger Proportion of Overlap Enhances Performance:** An increase in the proportion of overlap generally leads to improved performance, as shown in Figure 3 from left to right. A larger overlap results in more data samples, allowing the model to capture subtle variations in the data and enhance generalizability. Conversely, a smaller proportion of overlap yields fewer samples, which means that each incorrect classification has a proportionally larger impact on performance. It is noteworthy that the benefits of increasing the proportion of overlap are more pronounced when the window length below 40, suggesting that a window length of 40 might be optimal for emotion recognition.

Additionally, we compare the performance of our DSTER model across different emotions. Based on previous observations that the model performs optimally with a window length of 40 and a 90% proportion of overlap, we present results for these two settings: Table 4 details the F1

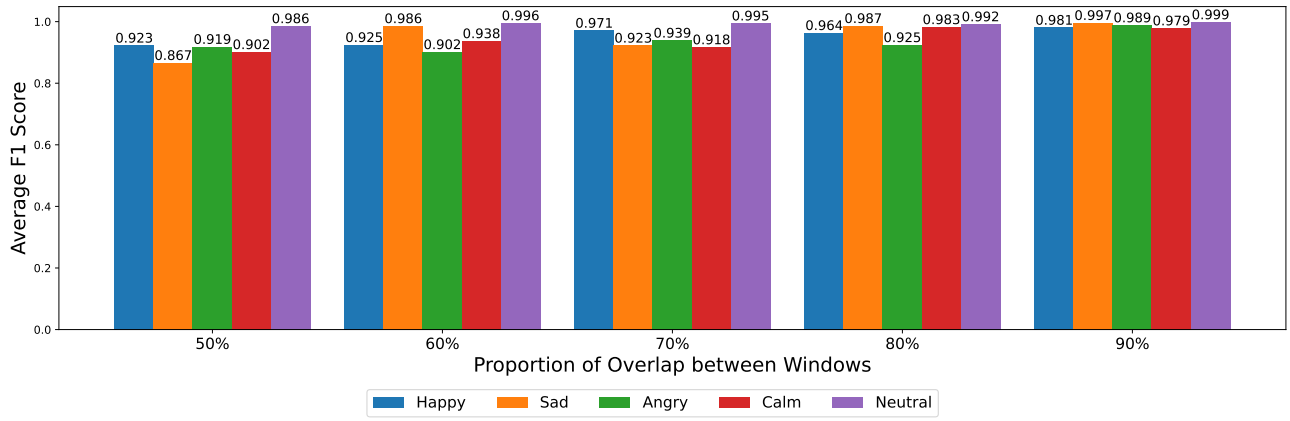

Figure 4. With window length of 40, the F1 score of DSTER model for each emotion at different proportions of overlap between samples.

scores for each emotion at different window lengths with a 90% overlap, while Figure 4 illustrates the F1 scores for each emotion at various overlap proportions with a window length of 40.

Window Length	Happy	Sad	Angry	Calm	Neutral
10	0.940	0.909	0.888	0.922	0.978
20	0.965	0.942	0.940	0.982	0.989
30	0.940	0.920	0.921	0.974	0.986
40	0.981	0.997	0.989	0.979	0.999
50	0.947	0.942	0.930	0.928	0.992
60	0.936	0.933	0.980	0.921	0.993

Table 4. With 90% proportion of overlap between samples, the F1 score of DSTER model for each emotion at different window length.

Consistent and Robust ER Performance Across All Emotions: The results demonstrate that our DSTER model achieves high and consistent performance across all emotions. Notably, it consistently excels at recognizing "Neutral," likely due to the significantly larger number of samples for "Neutral" compared to other emotions, as shown in Table 2. However, there are instances where its performance on certain emotions is slightly lower than on others. For example, with a window length of 40 and a 50% overlap, the model achieved an F1 score of only 0.867 for "Sad", which is 0.119 lower than the F1 scores for "Neutral". Despite these variations, the DSTER model does not consistently underperform on any particular emotion, suggesting that the observed performance differences are likely due to the sampling strategies or random participant partition.

5. Conclusion and Discussion

In this paper, we introduced the Dual-Stream Transformer-based Emotion Recognition (DSTER) model, which utilizes keystroke dynamics for emotion recognition. Our model incorporates a dual-stream architecture that effectively leverages both temporal-over-channel and channel-over-temporal information through a sophisticated yet effective combination of multi-head self-attention mechanisms, KDER Multi-LSTM blocks, and KDER Multi-CNN blocks. Dense vector embeddings for keycode information and Gaussian range encoding (GRE) further enhance the transformer's ability to process sequential and contextual information.

Experiments conducted on a widely recognized fixed-text keystroke dataset demonstrated that the DSTER model significantly outperforms existing baseline models, achieving 99.44% accuracy and an F1 score of 0.989 across various emotional states. Specifically, it showed an improvement in accuracy ranging from 22.96% to 66.04% over the baseline models from the past three years and maintained consistently high performance across all five tested emotions. Further analysis indicated that the DSTER model performs optimally with longer sample sizes and higher proportions of overlap, which allows the model to learn from subtle variations in the data, thereby enhancing its generalizability.

Despite the DSTER model's superior performance, there are areas for further improvement. Currently, our evaluation is limited to a single fixed-text dataset. Future work will involve assessing the model on free-text datasets, where the variable length of text samples presents a new challenge. A potential solution is to employ zero-padding to standardize sample lengths; however, the impact of such padding on model performance requires thorough investigation. Additionally, with the increasing ubiquity of mobile devices such as smartphones and tablets, extending our model to these platforms could significantly enhance human-computer interaction by deploying emotion recognition capabilities that dynamically adapt to users' needs.

References

[1] W. W. L. Z. Z. C. Bing Li, Wei Cui and M. Wu. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[2] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.

[3] H.-C. Chang, J. Li, C.-S. Wu, and M. Stamp. Machine learning and deep learning for fixed-text keystroke dynamics. In *Artificial Intelligence for Cybersecurity*, pages 309–329. Springer, 2022.

[4] D. Christian, M. Reynard, W. Astuti, G. Suharjanto, and R. Wulandari. Emotion recognition based on facial expression identification using deep learning algorithm for automation music healing application. In *E3S Web of Conferences*, volume 517, page 01002. EDP Sciences, 2024.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[6] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

[7] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. A. Aziz. Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. *Sensors*, 21(15):5015, 2021.

[8] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor. Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, 115:101–106, 2018.

[9] P. Kasprowski, Z. Borowska, and K. Harezlak. Biometric identification based on keystroke dynamics. *Sensors*, 22(9):3158, 2022.

[10] Y. Khaireddin and Z. Chen. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint arXiv:2105.03588*, 2021.

[11] A. Kołakowska. Recognizing emotions on the basis of keystroke dynamics. In *2015 8th International Conference on Human System Interaction (HSI)*, pages 291–297. IEEE, 2015.

[12] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15:99–117, 2012.

[13] P. Kumar and B. Raman. A bert based dual-channel explainable text emotion recognition system. *Neural Networks*, 150:392–407, 2022.

[14] M. S. Lee, Y. K. Lee, D. S. Pae, M. T. Lim, D. W. Kim, and T. K. Kang. Fast emotion recognition based on single pulse ppg signal with convolutional neural network. *Applied Sciences*, 9(16):3355, 2019.

[15] H. Locklear, S. Govindarajan, Z. Sitová, A. Goodkind, D. G. Brizan, A. Rosenberg, V. V. Phoha, P. Gasti, and K. S. Balagani. Continuous authentication with cognition-centric text production and revision features. In *Ieee international joint conference on biometrics*, pages 1–8. IEEE, 2014.

[16] A. Maalej and I. Kallel. Does keystroke dynamics tell us about emotions? a systematic literature review and dataset construction. In *2020 16th International Conference on Intelligent Environments (IE)*, pages 60–67. IEEE, 2020.

[17] A. Maalej and I. Kallel. Emosurv: A typing biometric (keystroke dynamics) dataset with emotion labels created using computer keyboards, 2020.

[18] A. Maalej, I. Kallel, and J. J. Sanchez Medina. Investigating keystroke dynamics and their relevance for real-time emotion recognition. *Available at SSRN 4250964*, 2023.

[19] S. Marrone, C. Sansone, et al. Identifying users' emotional states through keystroke dynamics. In *DeLTA*, pages 207–214, 2022.

[20] A. Mikuckas, I. Mikuckiene, A. Venckauskas, E. Kazanavicius, R. Lukas, and I. Plauska. Emotion recognition in human computer interaction systems. *Elektronika ir Elektrotechnika*, 20(10):51–56, 2014.

[21] Y. Qi, W. Jia, and S. Gao. Emotion recognition based on piezoelectric keystroke dynamics and machine learning. In *2021 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS)*, pages 1–4. IEEE, 2021.

[22] P. Sarkar and A. Etemad. Self-supervised learning for ecg-based emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3217–3221. IEEE, 2020.

[23] D. Senarath, S. Tharinda, M. Vishvajith, S. Rasnayaka, S. Wickramanayake, and D. Meedeniya. Behaveformer: A framework with spatio-temporal dual attention transformers for imu-enhanced keystroke dynamics. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2023.

[24] G. Stragapede, P. Delgado-Santos, R. Tolosana, R. Vera-Rodriguez, R. Guest, and A. Morales. Typeformer: Transformers for mobile keystroke biometrics. *arXiv preprint arXiv:2212.13075*, 2022.

[25] G. Stragapede, P. Delgado-Santos, R. Tolosana, R. Vera-Rodriguez, R. Guest, and A. Morales. Mobile keystroke biometrics using transformers. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2023.

[26] M. Trojahn, F. Arndt, M. Weinmann, and F. Ortmeier. Emotion recognition through keystroke dynamics on touchscreen keyboards. In *International Conference on Enterprise Information Systems*, volume 2, pages 31–37. SCITEPRESS, 2013.

[27] O. Valentin, A. Lehmann, D. Nguyen, and S. Paquette. Integrating emotion perception in rehabilitation programs for cochlear implant users: A call for a more comprehensive approach. *Journal of Speech, Language, and Hearing Research*, pages 1–8, 2024.

[28] A. Velichko and Y. Izotov. Emotions recognizing using lognnet neural network and keystroke dynamics dataset. In *AIP Conference Proceedings*, volume 2812. AIP Publishing, 2023.

[29] M. Xu, F. Zhang, and S. U. Khan. Improve accuracy of speech emotion recognition with attention head fusion. In

2020 10th annual computing and communication workshop and conference (CCWC), pages 1058–1064. IEEE, 2020.

[30] K. Yang, C. Wang, Y. Gu, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing*, 14(2):1082–1097, 2021.

[31] L. Yang and S.-F. Qin. A review of emotion recognition methods from keystroke, mouse, and touchscreen dynamics. *Ieee Access*, 9:162197–162213, 2021.