Frequency-Aware Gaze-based Authentication

Jonas Nasimzada
Department of Computer Science
University Stuttgart
Stuttgart, 70569
st171793@stud.uni-stuttgart.de

Álvaro González Tabernero
University Stuttgart
Stuttgart, 70569
st195627@stud.uni-stuttgart.de

Louis Beaudoing
University of Stuttgart
Stuttgart, 70569
st195304@stud.uni-stuttgart.de

Pranav Abraham Mathews
University of Stuttgart
Stuttgart, 70569
st191700@stud.uni-stuttgart.de

Abstract

In this work, three deep learning models — the original DenseNet model from the Eye Know You Too (EKYT) [1] study, a pre-trained Vision Transformer [2], and ResNet-101 [3] — are evaluated in order to explore gaze-based user authentication. Using data from the GazeBase dataset [4], the models were evaluated on a wide range of eye movement activities, including saccades, fixation, reading, and video viewing. DenseNet excelled with an Equal Error Rate (EER) of 3.24%. It performed exceptionally well on structured tasks but poorly on more dynamic ones. The Vision Transformer performed well in reading tasks by capturing long-term dependencies. However, it underperformed in limited and random tasks, resulting in an EER of 8.4%. ResNet-101 showed good performance on complicated tasks like random saccades and video viewing, but it was not generalizable, as seen by its EER of 10.49%. Over time, test-retest analysis showed that contextual influences and behavioural changes led to a decrease in authentication accuracy. The report points up a number of drawbacks, such as the short dataset size, information loss in specific tasks, and the computational requirements of sophisticated models. To address these limitations, the study highlights the importance of hybrid architectures and data augmentation techniques in improving the reliability and scalability of real-world biometric applications.

1 Introduction

Having to input the password to log into any device or application is becoming rarer by the day. Almost no portable device has been produced, independent of the category, that does not integrate some kind of biometric feature. This can come in the form of fingerprint scanner; modern solutions install some below the display, iris recognition, or face recognition. Technologies such as these have been readily available and regularly used for the better part of two decades. However, with the rapid development and increasing popularity in the realm of AR/VR products in recent times, devices in which optical user input is the main source of information, rather than an external keyboard or some sort of controller, point to the need of an update to biometric user authentication mechanisms.

Biometric features can be broken down into two distinct categories: physical and behavioural. Physical biometric features refer to concrete measures or characteristics of a person's body. These can range from fingerprints, irises, or facial factions to height or hair colour. On the other hand, behavioural biometric features are those intrinsic to one's way of doing a certain task or activity. For instance, walking patterns, voice intonation, or a simple signature can be classified as behavioural

biometric features. Their data collection and analysis are far less intrusive than their physical counterpart, even though physical features are inherent to everyone and tend to be more distinctive and permanent.

Eye movement biometrics (EMB) is a fairly recent behavioural biometric feature and the backbone of our project. Due to the complexity derived from the different muscles that take control of eye movement and the equally complex neurological signals and processes that go hand in hand with the actual physical movement, paired with differing biological features across different people, EMB can be considered an especially robust biometric feature. These qualities, combined with the growing use of eye tracking sensors in AR/VR devices, suggest that EMB could become a key security method in these systems.

Despite its potential, existing EMB models still struggle to meet the performance levels required for real-world use. Standards such as those established by the FIDO Alliance [5] recommend maintaining a false rejection rate of no more than 5% and a false acceptance rate of 1 in 10,000. Meeting these goals requires improvements in both the way models are designed, and the techniques used to handle eye movement data.

Our research builds on the Eye Know You Too (EKYT) [1] study but takes a different direction in terms of the model and data approach. We focus on a more modern architecture aimed at reducing complexity while improving performance. We also thoroughly test our model across a variety of tasks and recording rounds to better evaluate its effectiveness and ability to handle new scenarios.

2 Related Works

In recent years, machine and deep learning models have emerged as a cornerstone for processing biometric data, particularly eye movement signals. Convolutional Neural Networks (CNNs) have gained widespread adoption in image and time series analysis due to their remarkable ability to automatically learn meaningful features without the need for manual extraction. Pioneering works like AlexNet [6] and VGGNet [7] laid the foundation for CNN-based approaches in diverse fields. Nevertheless, in domains such as eye movement biometrics, conventional CNN architectures often encounter limitations when applied to real-world scenarios.

ResNet [3], introduced to address the issue of vanishing gradients in deep networks, introduced skip connections, which enhance the network's information flow. This innovation enables the training of significantly deeper networks compared to previous models. Research has shown that ResNet models can improve performance in both image recognition and time-series tasks, making them a strong contender for EMB.

Our work also dives into the application of transformer architectures. Unlike CNNs, vision transformers rely on self-attention mechanisms to capture long-term dependencies within a sequence. While Vision Transformers [2] have achieved remarkable success in natural language processing and image analysis, their potential in biometric time series analysis is still being explored. A significant challenge is that transformers typically necessitate substantial datasets to fully harness their modelling capabilities, which can pose a limitation in EMB, where datasets are often limited in size. Nevertheless, transformers demonstrate promise in capturing intricate temporal patterns in eye movement data.

Deep learning models designed for eye movement biometrics have undergone remarkable advancements over the past decade. Early models employed statistical approaches, such as the STAR model [8], which extracted hand-crafted features based on physiological eye events. In contrast, recent studies have embraced end-to-end learning, enabling neural networks to directly learn features from raw eye-tracking data. For instance, the DeepEyedentification-Live (DEL) [9] model employs separate subnetworks to analyse various types of eye movements, while the original Eye Know You (EKY) model utilized dilated convolutions to achieve robust results with reduced computational complexity.

While these models have made progress, they often lack modern architectural innovations like ResNet or transformers. Furthermore, many of them rely on relatively short input sequences (for instance, 1-second windows), which may hinder their ability to identify longer-term patterns. To address these shortcomings, our study integrates both ResNet and transformer architectures, while also investigating longer sequence lengths to enhance both feature extraction and model performance.

In essence, our research builds upon existing models such as EKY and DEL by incorporating modern, cutting-edge neural network architectures. Our objective is to assess the performance of these models under diverse conditions, including varying sampling rates, test-retest intervals, and compromised signal quality. By doing so, we strive to bridge the gap between theoretical concepts and practical, real-world applications, ultimately bringing EMB closer to its intended use.

3 Methodology

3.1 Research Design

Our research methodology combines replicability, comparability, and quantification. Replicability involves building upon an existing study, testing different models using the same dataset as the EKYT paper [1]. Comparability aims to evaluate multiple models and identify the most effective one. Quantification is achieved using various numerical performance metrics to assess the efficiency of our model.

In our implementation, we employ the Adam optimizer along with a OneCycleLR learning rate scheduler, which dynamically adjusts the learning rate throughout the training process. The model is trained for 100 epochs, with a maximum learning rate of 0.01. The scheduler controls the learning rate using decay factors, with `div_factor = 100.0` and `final_div_factor = 1000.0`.

3.2 Data Collection

The dataset used in this study originates from the EKYT paper and was obtained from Dataverse, hosted in a Texas State University repository. It comprises 322 college participants, each recorded monocularly (using only their left eye) at a frequency of 1000 Hz using an EyeLink 1000 eye tracker. The data was captured over nine rounds of recordings. Notably, the dataset was already pre-processed, which enabled us to concentrate on model experimentation without the need for additional data-cleaning steps.

3.3 Tools and Frameworks

To collaboratively develop and test the models, we utilized PyCharm, which facilitated real-time teamwork within a shared Integrated Development Environment (IDE). The project was implemented in Python, with PyTorch and PyTorch Lightning serving as the primary deep-learning frameworks. To enhance computational efficiency and speed up training, model training, testing and validations were conducted on a laptop GPU.

3.4 Experimental Procedure

3.4.1 Data Transformation

We adopted an analysis approach based on spectrograms. A spectrogram is a visual representation of the spectrum of frequencies in a signal as it varies over time. It is generated by applying a Short-Time Fourier Transform (STFT), which breaks down the signal into time-frequency components, enabling a detailed analysis of both short-term and long-term variations.

In our use case, this transformation offers several benefits. Eye movement signals, particularly in tasks such as Random Saccades and video viewing, exhibit dynamic and frequency-based patterns that can be challenging to capture using raw time-series data. By converting the data into spectrogram representations (Fig. 1), we highlight key temporal and frequency features, making it easier for deep learning models to identify distinguishing characteristics. This is particularly relevant for models such as DenseNet and ResNet [3], which excel at extracting hierarchical features from image-like input data.

Furthermore, spectrograms enhance task-specific performance by providing clearer patterns for structured tasks such as Reading (TEX), where consistent eye movement frequencies are crucial for accurate authentication. Consequently, spectrograms serve as an effective preprocessing step, facilitating robust feature extraction across diverse tasks and enhancing model performance.

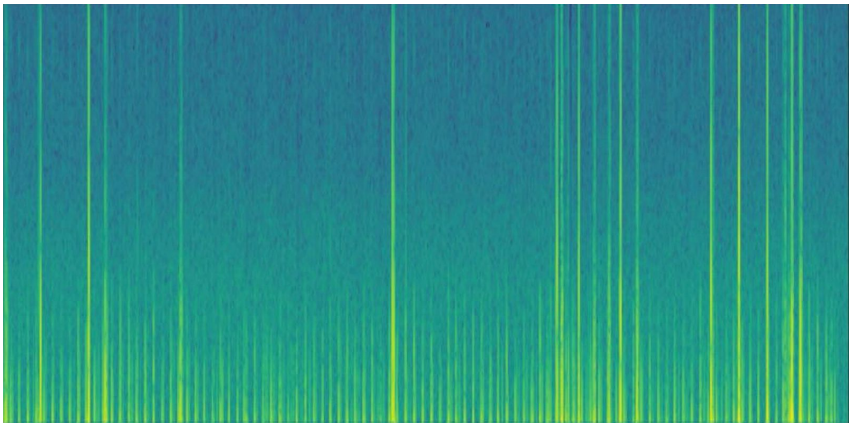

Figure 1: Spectrogram representation of eye movement data

3.4.2 Initial Model Training

Initially, we employed a Vision Transformer network [2] trained on spectrograms, with a specific focus on audio-based analysis. Furthermore, some team members investigated suitable algorithms specifically designed for spectrogram analysis. The assumption was that the features that were learnt for audio data could be transferred and might provide valuable insights into eye-movement frequency spectrograms as well.

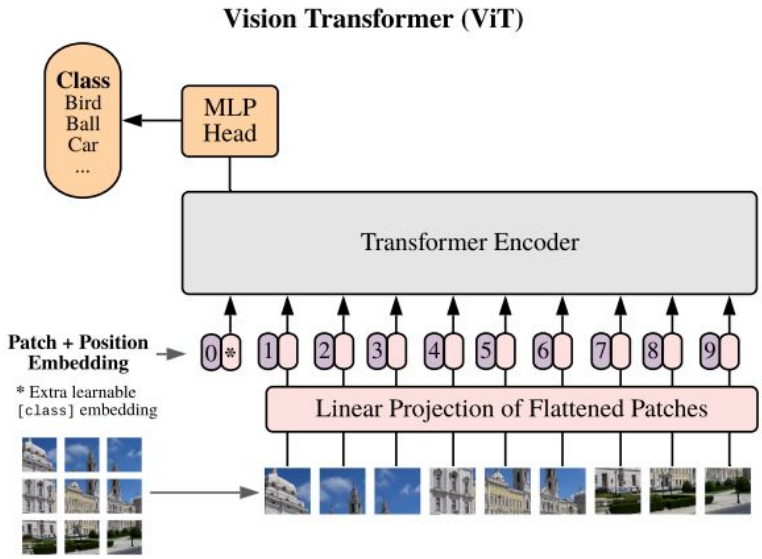

Figure 2: Vision Transformer Architecture [10]

3.4.3 Model Optimization

Upon receiving the initial results, we made the decision to investigate alternative architectures, including ResNet-50 and ResNet-101, to attempt enhancing the performance of the while keeping model complexity simple enough (Fig. 3) to accommodate our computational resources. Subsequently, we conducted further training and evaluation to compare these models against the baseline model.

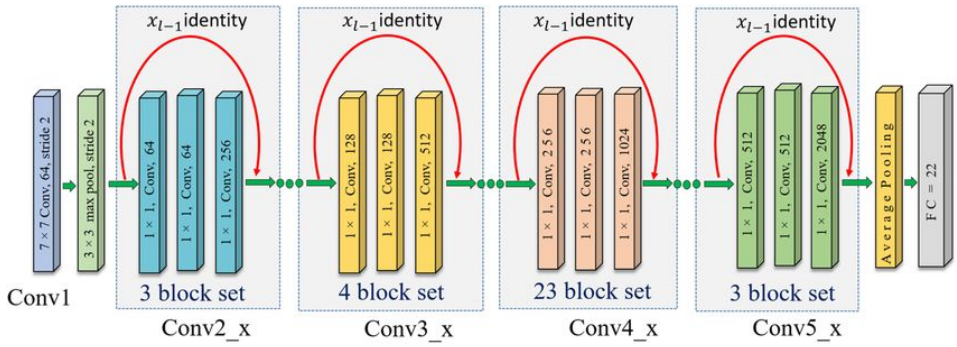

Figure 3: ResNet Architecture [11]

3.5 Evaluation

To evaluate the efficacy of diverse models, we employed the Equal Error Rate (EER) metric. EER is widely utilized in biometric and verification scenarios, signifying the juncture where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) attain parity. A lower EER denotes superior model performance. Measurement of EER requires registration of data that is used as the reference for computing the pairwise similarity against the data required for authentication. Pairwise cosine similarities are computed between the enrolment and authentication data, and the resulting similarity scores and then used as input to a receiver operating characteristic (ROC) curve to obtain the EER values.

Furthermore, the concept of decidability (d') provides additional insight into model performance by quantifying the separability between the distributions of genuine and impostor scores. Decidability is defined by the formula:

$$d' = \frac{|\mu_G - \mu_I|}{\sqrt{\frac{\sigma_G^2 + \sigma_I^2}{2}}} \tag{1}$$

In this study, μ_G and μ_I denote the means of the genuine and impostor score distributions, respectively, while σ_G and σ_I denote their standard deviations. Higher values of d' indicate greater separation between the two distributions, signifying a clearer distinction between legitimate users and impostors. Decidability has the benefit of being a threshold-free metric compared to EER, since there might be points on the Receiver Operating Curve (ROC) where there might not be a similarity threshold where FAR and FRR meet. In such cases, decidability can prove to be a more useful metric to determine the performance, although a threshold would still have to be set if the results of this study were to be used in biometric settings.

The evaluation pipeline for computing the equal error rate and the decidability metrics are adapted from the baseline study.

In this study, both EER and decidability is used as the evaluation metric to compare the DenseNet, Vision Transformer, and ResNet models. By examining these metrics, we gained a comprehensive view of the ability of each model to provide accurate and robust gaze-based authentication.

4 Experiments

4.1 Dataset

Seven separate tasks were employed to capture various eye movement behaviours. The Horizontal Saccades (HSS) task presented a target that moved horizontally across the screen at regular intervals, prompting systematic saccadic movements. In contrast, the Random Saccades (RAN) task introduced unpredictability by having the target change position randomly at regular intervals.

The Fixation (FXS) task required participants to maintain a steady gaze on a static target for 15 seconds. The Reading (TEX) task involved reading a passage for 60 seconds, capturing natural

reading-related eye movements. Two video-viewing tasks were also included: Video Viewing 1 (VD1) displayed the first minute of a movie trailer to record eye movements in a cinematic context, while Video Viewing 2 (VD2) presented the subsequent minute of the trailer.

The Balura Game (BLG) task involved a gaze-driven interaction where participants were tasked with eliminating red balls by fixating on them. The objective was to clear all targets as quickly as possible. Due to its variable task duration, the BLG task was excluded from the original EKYT [1] study but was included in this investigation for further evaluation of the model's performance on out-of-distribution data.

The above tasks were exposed to the participants of the dataset in 9 rounds of recording that were captured over a span of 37 months, and within each round of recording, two sessions were undertaken spaced apart by 30 minutes each. It can also be noted that each subsequent round of recording had only a subset of participants from the immediate previous rounds, with only 322 participants present for all rounds.

4.2 Analysis

We performed two distinct performance analyses on three different models using two metrics: the EER and decidability scores. The first model tested was the original paper's DenseNet CNN, followed by the pre-trained Vision Transformer [2], and finally the ResNet-101 [3]. The tasks were twofold: firstly, we measured task performance across all seven tasks in the dataset, and then we measured each model's performance on a test-retest framework, where the focus was on identifying the best-performing model over time. The results are presented in Figure 4–9 below.

The primary distinction between the two analyses lies in their enrolment and authentication test methodologies. The original paper's approach, the first test we would attempt, collected enrolment data from the initial session of the first round, while authentication data used for evaluation came from the second session of that same first round. This test design prioritized short-term model performance, with only thirty minutes separating the two sessions. In contrast, our second test design addressed this issue by testing any enrolment data from a specific first session against any authentication data from a corresponding second session, regardless of the round. The results are presented in Figure 10–15 below.

4.3 Results

It should be noted that, to properly convey the effects of variations in the tasks and test-retests, we applied a scale that is respective to each experiment. Using the same range within the scale would have resulted in a shift of focus of the presented results to the performance disparity between the original study and our own. Due to the quantity of data and presentation constraints, more detailed and tabulated results can be found in the appendix.

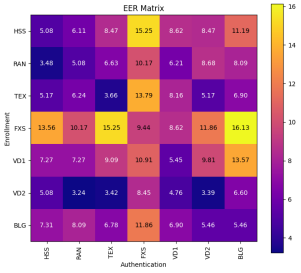

Figure 4: EKYT's DenseNet EER Results

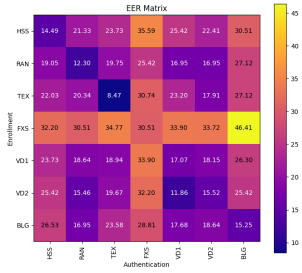

Figure 5: Vision Transformer EER Results

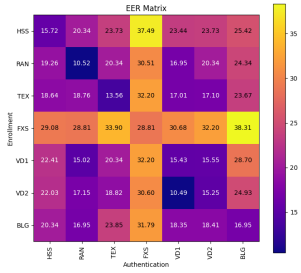

Figure 6: ResNet-101 EER Results

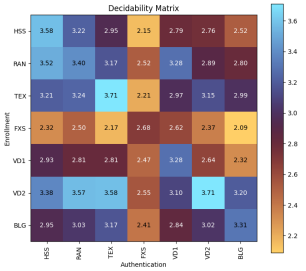

Figure 7: EKYT's DenseNet Decidability

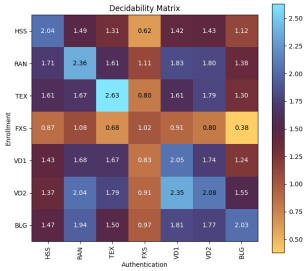

Figure 8: Vision Transformer Decidability

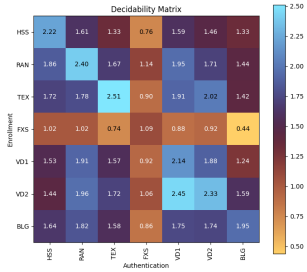

Figure 9: ResNet-101 Decidability

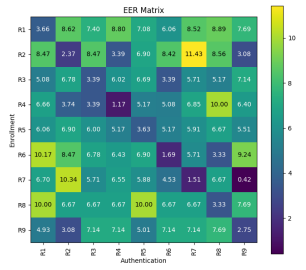

Figure 10: EKYT's DenseNet EER Results

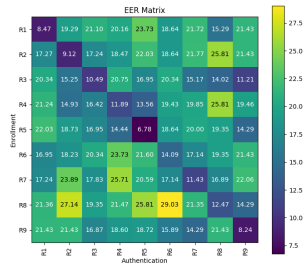

Figure 11: Vision Transformer EER Results

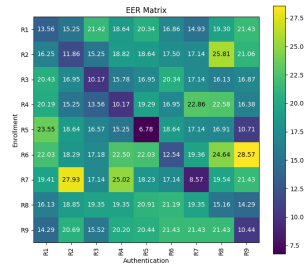

Figure 12: ResNet-101 EER Results

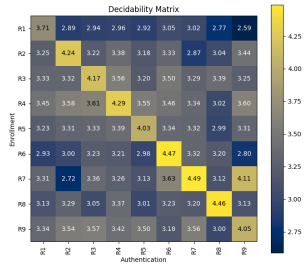

Figure 13: EKYT's DenseNet Decidability

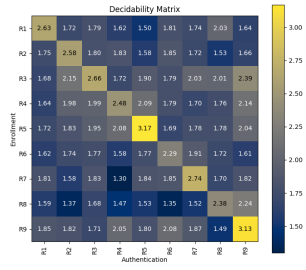

Figure 14: Vision Transformer Decidability

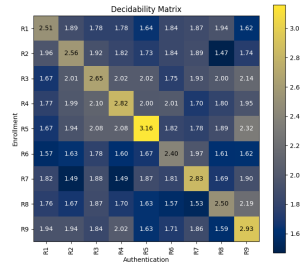

Figure 15: ResNet-101 Decidability

5 Discussion

5.1 Effect of Task on Authentication Accuracy

In general, the DenseNet architecture of the original paper remains the most effective model, achieving an Equal Error Rate (EER) of 3.24% (Fig. 4), the lowest in all test scenarios. This performance underscores its strength in tasks with consistent and minimally variable eye movements, such as Reading (TEX) and Horizontal Saccades (HSS). However, despite its robust overall accuracy, the model encounters significant limitations when processing more unpredictable eye patterns, particularly in tasks like Random Saccades (RAN), where eye movements react rapidly to random stimuli. Additionally, DenseNet struggles with tasks involving complex or highly dynamic visual content, such as video viewing.

In second place, the pre-trained Vision Transformer [2] achieved an EER of 8.4% (Fig. 5), demonstrating weaker general performance compared to DenseNet. One of the primary reasons for this lower accuracy is the difficulty transformers encounter when dealing with constrained or erratic tasks like Horizontal Saccades and Random Saccades. These tasks do not provide the long-term temporal relationships that transformers excel at capturing. Furthermore, due to the constraints of our testing and dataset size, our Transformer started to show signs of small overfitting. However, the Vision

Transformer exhibits a distinct advantage in tasks such as reading, where structured and sequential eye movement patterns align with its ability to model long-term dependencies. This strength enables it to outperform other models in these specific tasks despite its overall higher EER.

The ResNet-101 model ranked third, recording an EER of 10.49% (Fig. 6). Although this is significantly higher than that of the DenseNet, the ResNet demonstrates promising capabilities when tasked with more complex or feature-rich scenarios, including Random Saccades and video viewing. Its architecture appears to be better suited to capture subtle frequency-based and temporal features inherent in these tasks. However, its higher error rate suggests that, while capable of handling nuanced patterns, it may lack the generalization and parameter efficiency observed in DenseNet.

The inherent characteristics of the tasks themselves significantly influence the performance of the models. Structured tasks with minimal variability, such as Reading (TEX) and Fixation (FXS), provide consistent input, enabling models to extract reliable features. Reading tasks, in particular, consistently yielded robust results for both DenseNet and the Vision Transformer due to their sequential and patterned nature. In contrast, fixation tasks performed poorly because the limited movement data hindered the models' ability to distinguish between users. Tasks like Random Saccades and video viewing presented the greatest challenge due to their high unpredictability, with models struggling to maintain high accuracy across these scenarios. While ResNet exhibited some capacity to manage these tasks, its performance remained suboptimal compared to DenseNet in other domains.

The same trends can also be observed within the decidability score matrices (Fig. 4–6) although it can be observed that the performance disparity between the original work and our study is much less evident. Therefore, decidability could potentially serve as a better metric in evaluation scenarios and, upon the adoption of a threshold, could also be considered for biometric authentication.

5.2 Effect of Test-Retest on Authentication Accuracy

Another crucial aspect of the study was the evaluation of test-retest performance using data from multiple rounds of the GazeBase dataset [4]. Over time, the accuracy of the authentication declined, which can be attributed to various factors, such as participant fatigue, time of day, and evolving reading or behavioural patterns. DenseNet maintained relatively stable performance over short-term time intervals but exhibited greater fluctuations over extended periods. The Vision Transformer exhibited inconsistent performance across rounds, probably due to its sensitivity to changing task patterns over time. ResNet, on the contrary, demonstrated a more balanced performance in all rounds but remained constrained by its higher overall error rate.

These observations also reveal significant limitations of the models and experimental design. Firstly, the issue of information loss is evident, particularly in tasks that do not provide sufficient movement data or when task transitions reduce the continuity of eye-tracking signals. Secondly, the relatively small size of the GazeBase dataset limits the generalizability of complex models, such as transformers, increasing the risk of overfitting. Overfitting is particularly relevant for DenseNet, which may have achieved its high accuracy due to over-adaptation to the specific eye movement patterns in the dataset. Furthermore, the computational demands of certain models, particularly the Vision Transformer, present another challenge. Transformers require extensive computational resources and larger datasets to fully realize their potential, making them less practical for smaller biometric datasets or real-world applications without further optimization.

As for the test-retest experiments, similar patterns can also be observed within the decidability score matrices (Fig. 13–15).

6 Conclusion

This study evaluated the performance of gaze-based authentication models, including DenseNet from the original EKYT paper [1], a pre-trained Vision Transformer [2], and ResNet-101 [3]. DenseNet achieved the lowest error rate (EER) of 3.24%, outperforming the other models on structured tasks like Reading (TEX) and Horizontal Saccades (HSS). However, it struggled with unpredictable tasks like Random Saccades (RAN) and video viewing, suggesting limited generalization capabilities.

The Vision Transformer, with an EER of 8.4%, excelled in reading tasks by capturing long-term dependencies. However, it performed poorly on tasks requiring constrained or erratic movements.

ResNet-101 achieved an EER of 10.49%, excelling on complex tasks like video viewing but exhibiting limited generalization.

Test-retest analysis using the GazeBase dataset revealed performance degradation over time due to fatigue and behavioral changes. While DenseNet maintained stable short-term performance, both it and the Vision Transformer showed fluctuations over longer intervals. ResNet demonstrated more balanced performance despite higher error rates.

Potential limitations included the small dataset size, task-specific information loss, and substantial computational demands, particularly for the Vision Transformer. Future research should explore hybrid models, data augmentation techniques, and larger datasets to enhance both accuracy and robustness for practical applications.

7 References

[1] D. Lohr and O. V. Komogortsev, "Eye Know You Too: Toward viable end-to-end eye movement biometrics for user authentication," *IEEE Transactions on Information Forensics and Security*, vol. 17, Aug. 2022, doi: 10.1109/TIFS.2022.3201369.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[4] H. Griffith, D. Lohr, E. Abdulin, and O. Komogortsev, "Gazebase: A large-scale multi-stimulus longitudinal eye movement dataset," *Scientific Data*, vol. 8, no. 1, p. 184, Jul. 2021.

[5] S. Schuckers, G. Cannon, and N. Tekampe, "FIDO biometrics requirements," Apr. 2021. [Online]. Available: https://fidoalliance.org/specs/biometric/requirements/.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25. Red Hook, NY, USA: Curran Associates, 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, vol. 1409.1556, 2015.

[8] L. Friedman, M. S. Nixon, and O. V. Komogortsev, "Method to assess the temporal persistence of potential biometric features: Application to oculomotor gait face and brain structure databases," *PLoS One*, vol. 12, no. 6, pp. 1–42, Jun. 2017.

[9] S. Makowski, P. Prasse, D. R. Reich, D. Krakowczyk, L. A. Jäger, and T. Scheffer, "DeepEyedentificationLive: Oculomotoric biometric identification and presentation-attack detection using deep neural networks," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 506–518, Oct. 2021.

[10] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022. [Online]. Available: https://docs.nvidia.com/nemo-framework/user-

guide/24.07/nemotoolkit/vision/vit.html.

[11]V. Tiwari, R. C. Joshi, and M. K. Dutta, "Deep neural network for multi-class classification of medicinal plant leaves," *ResearchGate*, 2022. [Online]. Available: https://www.researchgate.net/publication/360641512_Deep_neural_network_for_multi-class_classification_of_medicinal_plant_leaves

8 Appendix

All of our code can be found at this repository

Table 1: Effect of Task on Authentication Equal Error Rate (EER)

	Authentication						
Enrollment	HSS	RAN	TEX	FXS	VD1	VD2	BLG
DenseNet (EKYT)							
HSS	5.08	6.11	8.47	15.25	8.62	8.47	11.19
RAN	3.48	5.08	6.63	10.17	6.21	8.68	8.09
TEX	5.17	6.24	3.66	13.79	8.16	5.17	6.90
FXS	13.56	10.17	15.25	9.44	8.62	11.86	16.13
VD1	7.27	7.27	9.09	10.91	5.45	9.81	13.57
VD2	5.08	*3.24	3.42	8.45	4.76	3.39	6.60
BLG	7.31	8.09	6.78	11.86	6.90	5.46	5.46
Vision Transformer							
HSS	14.49	21.33	23.73	35.59	25.42	22.41	30.51
RAN	19.05	12.30	19.75	25.42	16.95	16.95	27.12
TEX	22.03	20.34	*8.47	30.74	23.20	17.91	27.12
FXS	32.20	30.51	34.77	30.51	33.90	33.72	46.41
VD1	23.73	18.64	18.94	33.90	17.07	18.15	26.30
VD2	25.42	15.46	19.67	32.20	11.86	15.52	25.42
BLG	26.53	16.95	23.58	28.81	17.68	18.64	15.25
ResNet-101							
HSS	15.72	20.34	23.73	37.49	23.44	23.73	25.42
RAN	19.26	10.52	20.34	30.51	16.95	20.34	24.34
TEX	18.64	18.76	13.56	32.20	17.01	17.10	23.67
FXS	29.08	28.81	33.90	28.81	30.68	32.20	38.31
VD1	22.41	15.02	20.34	32.20	15.43	15.55	28.70
VD2	22.03	17.15	18.82	30.60	*10.49	15.25	24.93
BLG	20.34	16.95	23.85	31.79	18.35	18.41	16.95

*Highest authentication equal error rate (EER) achieved in task combinations.

Table 2: Effect of Task on Authentication Decidability (d')

	Authentication						
Enrollment	HSS	RAN	TEX	FXS	VD1	VD2	BLG
DenseNet (EKYT)							
HSS	3.58	3.22	2.95	2.15	2.79	2.76	2.52
RAN	3.52	3.40	3.17	2.52	3.28	2.89	2.80
TEX	3.21	3.24	*3.71	2.21	2.97	3.15	2.99
FXS	2.32	2.50	2.17	2.68	2.62	2.37	2.09
VD1	2.93	2.81	2.81	2.47	3.28	2.64	2.32
VD2	3.38	3.57	3.58	2.55	3.10	*3.71	3.20
BLG	2.95	3.03	3.17	2.41	2.84	3.02	3.31
Vision Transformer							
HSS	2.04	1.49	1.31	0.62	1.42	1.43	1.12
RAN	1.71	2.36	1.61	1.11	1.83	1.80	1.38
TEX	1.61	1.67	*2.63	0.80	1.61	1.79	1.30
FXS	0.87	1.08	0.68	1.02	0.91	0.80	0.38
VD1	1.43	1.68	1.67	0.83	2.05	1.74	1.24
VD2	1.37	2.04	1.79	0.91	2.35	2.08	1.55
BLG	1.47	1.94	1.50	0.97	1.81	1.77	2.03
ResNet-101							
HSS	2.22	1.61	1.33	0.76	1.59	1.46	1.33
RAN	1.86	2.40	1.67	1.14	1.95	1.71	1.44
TEX	1.72	1.78	*2.51	0.90	1.91	2.02	1.42
FXS	1.02	1.02	0.74	1.09	0.88	0.92	0.44
VD1	1.53	1.91	1.57	0.92	2.14	1.88	1.24
VD2	1.44	1.96	1.72	1.06	2.45	2.33	1.59
BLG	1.64	1.82	1.58	0.86	1.75	1.74	1.95

*Highest authentication decidability (d') achieved in task combinations.

Table 3: Effect of Test-Retest on Authentication Equal Error Rate (EER)

	Authentication								
Enrollment	R1	R2	R3	R4	R5	R6	R7	R8	R9
DenseNet (EKYT)									
R1	3.66	8.62	7.40	8.80	7.08	6.06	8.52	8.89	7.69
R2	8.47	2.37	8.47	3.39	6.90	8.42	11.43	8.56	3.08
R3	5.08	6.78	3.39	6.02	6.69	3.39	5.71	5.17	7.14
R4	6.66	3.74	3.39	1.17	5.17	5.08	6.85	10.00	6.40
R5	6.06	6.90	6.00	5.17	3.63	5.17	5.91	6.67	5.51
R6	10.17	8.47	6.78	4.43	6.90	1.69	5.71	3.33	9.24
R7	6.70	10.34	5.71	6.55	5.88	4.53	1.51	6.67	*0.42
R8	10.00	6.67	6.67	6.67	10.00	6.67	6.67	3.33	7.69
R9	4.93	3.08	7.14	7.14	5.01	7.14	7.14	7.69	2.75
Vision Transformer									
R1	8.47	19.29	21.10	20.16	23.73	18.64	21.72	15.29	21.43
R2	17.27	9.12	17.24	18.47	22.03	18.64	21.77	25.81	21.43
R3	20.34	15.25	10.49	20.75	16.95	20.34	15.17	14.02	11.21
R4	21.24	14.93	16.42	11.89	13.56	19.43	19.85	25.81	19.46
R5	22.03	18.73	16.95	14.44	*6.78	18.64	20.00	19.35	14.29
R6	16.95	18.23	20.34	23.73	21.60	14.09	17.14	19.35	21.43
R7	17.24	23.89	17.83	25.71	20.59	17.14	11.43	16.89	22.06
R8	21.36	27.14	19.35	21.47	25.81	29.03	21.35	12.47	14.29
R9	21.43	21.43	16.87	18.60	18.72	15.89	14.29	21.43	8.24
ResNet-101									
R1	13.56	15.25	21.42	18.64	20.34	16.86	14.93	19.30	21.43
R2	16.25	11.86	15.25	18.82	18.64	17.50	17.14	25.81	21.06
R3	20.43	16.95	10.17	15.78	16.95	20.34	17.14	16.13	16.87
R4	20.19	15.25	13.56	10.17	19.29	16.95	22.86	22.58	16.38
R5	23.55	18.64	16.57	15.25	*6.78	18.64	17.14	16.91	10.71
R6	22.03	18.29	17.18	22.50	22.03	12.54	19.36	24.64	28.57
R7	19.41	27.93	17.14	25.02	18.23	17.14	8.57	19.54	21.43
R8	16.13	18.85	19.35	19.35	20.91	21.19	19.35	15.16	14.29
R9	14.29	20.69	15.52	20.20	20.44	21.43	21.43	21.43	10.44

*Highest authentication equal error rate (EER) achieved in test-retest combinations.

Table 4: Effect of Test-Retest on Authentication Decidability (d')

	Authentication								
Enrollment	R1	R2	R3	R4	R5	R6	R7	R8	R9
DenseNet (EKYT)									
R1	3.71	2.89	2.94	2.96	2.92	3.05	3.02	2.77	2.59
R2	3.25	4.24	3.22	3.38	3.18	3.33	2.87	3.04	3.44
R3	3.33	3.32	4.17	3.56	3.20	3.50	3.29	3.39	3.25
R4	3.45	3.58	3.61	4.29	3.55	3.46	3.34	3.02	3.60
R5	3.23	3.31	3.33	3.39	4.03	3.34	3.32	2.99	3.31
R6	2.93	3.00	3.23	3.21	2.98	4.47	3.32	3.20	2.80
R7	3.31	2.72	3.36	3.26	3.13	3.63	*4.49	3.12	4.11
R8	3.13	3.29	3.05	3.37	3.01	3.23	3.20	4.46	3.13
R9	3.34	3.54	3.57	3.42	3.50	3.18	3.56	3.00	4.05
Vision Transformer									
R1	2.63	1.72	1.79	1.62	1.50	1.81	1.74	2.03	1.64
R2	1.75	2.58	1.80	1.83	1.58	1.85	1.72	1.53	1.66
R3	1.68	2.15	2.66	1.72	1.90	1.79	2.03	2.01	2.39
R4	1.64	1.98	1.99	2.48	2.09	1.79	1.70	1.76	2.14
R5	1.72	1.83	1.95	2.08	*3.17	1.69	1.78	1.78	2.04
R6	1.62	1.74	1.77	1.58	1.77	2.29	1.91	1.72	1.61
R7	1.81	1.58	1.83	1.30	1.84	1.85	2.74	1.70	1.82
R8	1.59	1.37	1.68	1.47	1.53	1.35	1.52	2.38	2.24
R9	1.85	1.82	1.71	2.05	1.80	2.08	1.87	1.49	3.13
ResNet-101									
R1	2.51	1.89	1.78	1.78	1.64	1.84	1.87	1.94	1.62
R2	1.96	2.56	1.92	1.82	1.73	1.84	1.89	1.47	1.74
R3	1.67	2.01	2.65	2.02	2.02	1.75	1.93	2.00	2.14
R4	1.77	1.99	2.10	2.82	2.00	2.01	1.70	1.80	1.95
R5	1.67	1.94	2.08	2.08	*3.16	1.82	1.78	1.89	2.32
R6	1.57	1.63	1.78	1.60	1.67	2.40	1.97	1.61	1.62
R7	1.82	1.49	1.88	1.49	1.87	1.81	2.83	1.69	1.90
R8	1.76	1.67	1.87	1.70	1.63	1.57	1.53	2.50	2.19
R9	1.94	1.94	1.84	2.02	1.63	1.71	1.86	1.59	2.93

*Highest authentication decidability (d') achieved in test-retest combinations.

Contribution Statement

The following table illustrates the contributions of all members in preparing this work.

Chapter Heading	**Jonas Nasimzada**	**Álvaro González Tabernero**	**Louis Beaudoing**	**Pranav Abraham Mathews**
Introduction & Related Work	25%	25%	25%	25%
Methodology & Experimental Setup	25%	25%	25%	25%
Results & Discussion	25%	25%	25%	25%
Conclusion	25%	25%	25%	25%
Final Editing & Proofreading	25%	25%	25%	25%