ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

«Αναγνώριση συναισθήματος σε παιχνίδια σοβαρού σκοπού για άτομα που πάσχουν από ψυχικές διαταραχές.»

Εξαμηνιαία Εργασία

στο μάθημα «**Τεχνολογίες Κινητής και Ηλεκτρονικής Υγείας**»

των φοιτητών

Ηλιόπουλος Γεώργιος, Α.Μ.: 03118815

Ράτσα Ηλία, Α.Μ.: 03118817

Σερλής Εμμανουήλ-Αναστάσιος, Α.Μ.: 03118125

Σπανός Νικόλαος, Α.Μ.: 03118822

Διδάσκοντες: Δ. Κουτσούρης, Κ. Νικήτα, Γ. Ματσόπουλος

Υπεύθυνος Άσκησης: Κώστας Μήτσης

Αθήνα, Ιούνιος 2022

Η σελίδα αυτή είναι σκόπιμα λευκή.

Περίληψη

Τα παιχνίδια σοβαρού σκοπού (serious games) έχουν λάβει αυξανόμενο ερευνητικό ενδιαφέρον τα τελευταία χρόνια, με αποτέλεσμα να αναγνωρίζονται ως ένα χρήσιμο εργαλείο στα χέρια των ειδικών υγείας. Ένας διαρκώς εξελισσόμενος τομέας εφαρμογών τους αποτελεί η ψυχική υγεία, με τα μέχρι στιγμής αποτελέσματα να είναι πολλά υποσχόμενα για την υποβοήθηση ατόμων με ψυχικές διαταραχές. Σε αυτό το πλαίσιο, σκοπός της εργασίας αποτελεί η υλοποίηση ενός σοβαρού παιχνιδιού για την εκμάθηση συναισθημάτων σε άτομα που αντιμετωπίζουν δυσκολία στην αναγνώριση τους. Πιο αναλυτικά, κατά την αρχική αλληλεπίδραση του παίκτη με το παιχνίδι, θα προσδιορισθεί η συναισθηματική κατάστασή του από ένα κατάλληλα σχεδιασμένο νευρωνικό δίκτυο αναγνώρισης συναισθημάτων σε πραγματικό χρόνο, το οποίο θα δέχεται ως είσοδο την φωνή του χρήστη. Στην συνέχεια, ο παίκτης θα εισέρχεται σε ένα παιχνίδι καρτών, το οποίο θα αποσκοπεί στην εκμάθηση διαφόρων συναισθηματικών καταστάσεων. Αυτό θα επιτυγχάνεται με το να του ζητείται να αναγνωρίσει το συναίσθημα που παρουσιάζεται πάνω στις κάρτες και θα γίνεται η συλλογή των δεδομένων της επιτυχίας. Επίσης, αποσκοπεί στο να δημιουργεί περιστατικά πάνω στα οποία βασίζεται το αναγνωριζόμενο συναίσθημα. Η εν λόγω υλοποίηση ευελπιστεί να αξιοποιήσει ευφυώς τα συναισθήματα του παίκτη, έτσι ώστε να αναδειχθεί η δυνατότητα εκμάθησης των ίδιων των συναισθημάτων του με μία ευχάριστη και διαδραστική μέθοδο. Ταυτόχρονα, στοχεύει και στην ευρύτερη μετάδοση των παιχνιδιών σοβαρού σκοπού(serious games) ως μία βιώσιμη και εφικτή λύση, στα πλαίσια εξατομίκευσης της κλινικής διαδικασίας, μέσω του προσδιορισμού ατομικών χαρακτηριστικών, όπως η ψυχολογική κατάσταση και η νοητική ικανότητα του ασθενούς.

Λέξεις Κλειδιά

serious game; emotion recognition; emotional learning; neural network; medical individualization.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

Η σελίδα αυτή είναι σκόπιμα λευκή.

1 *Εισαγωγή*

Η παρούσα εργασία έχει ως σκοπό να διερευνήσει την εφαρμογή νευρωνικών δικτύων και παιχνιδιών σοβαρού σκοπού στην ενίσχυση ατόμων με ψυχικές διαταραχές. Η εφαρμογή αποτελείται από ένα κομμάτι αναγνώρισης συναισθήματος μέσω ηχητικών δειγμάτων και την χρήση αυτής της αναγνώρισης δυναμικά σε ένα παιχνίδι σοβαρού σκοπού το οποίο αποσκοπεί στην ενίσχυση της συναισθηματικής νοημοσύνης των ατόμων με ψυχικές διαταραχές. Τα νευρωνικά δίκτυα αποτελούν έναν ανερχόμενο τομέα της επιστήμης υπολογιστών και εφαρμογές τους εμφανίζονται σε πολλούς τομείς της καθημερινότητας και της επιστήμης, με ανερχόμενες επεκτάσεις και στην ιατρική.

Από την βιβλιογραφία παρατηρήθηκε ότι υπάρχει μεγάλη ποικιλία τέτοιων νευρωνικών συστημάτων [1]. Οι μέθοδοι αναγνώρισης συναισθημάτων μέσω φωνής βασίστηκαν σε αυτές που χρησιμοποιήθηκαν στην αυτόματη αναγνώριση ομιλίας (Automatic Speech Recognition - ASR), δηλαδή HMMS, GMMs και SVMs [1]. Όμως, η αδυναμία αυτών των τεχνικών ήταν ότι οποιεσδήποτε αλλαγές στα χαρακτηριστικά απαιτούσαν συνήθως αναδιάρθρωση ολόκληρης της αρχιτεκτονικής της μεθόδου.

Ωστόσο, τα τελευταία χρόνια, έχουν αναπτυχθεί εργαλεία και διαδικασίες, πιο συγκεκριμένα δημιουργία και εξέλιξη των βαθιών νευρωνικών δικτύων (Deep Neural Network - DNN), των συνελικτικών νευρωνικών δικτύων (Convolutional Neural Network - CNN) και των επαναλαμβανόμενων νευρωνικών δικτύων (Recurrent Neural Networks - RNN), με τα οποία δεν υφίσταται, πλέον, τέτοιο πρόβλημα [1], [2], [3], [4]. Σύμφωνα με την υπάρχουσα βιβλιογραφία το πιο διαδεδομένο και επιτυχημένο μοντέλο είναι το δίκτυο μακράς βραχύχρονης μνήμης (Long Short-Term Memory - LSTM), το οποίο συνηθίζεται να χρησιμοποιεί σαν χαρακτηριστικό εισόδου τα Mel Frequency Cepstrum Coefficients (MFCC) [1], [2], [3], [4].

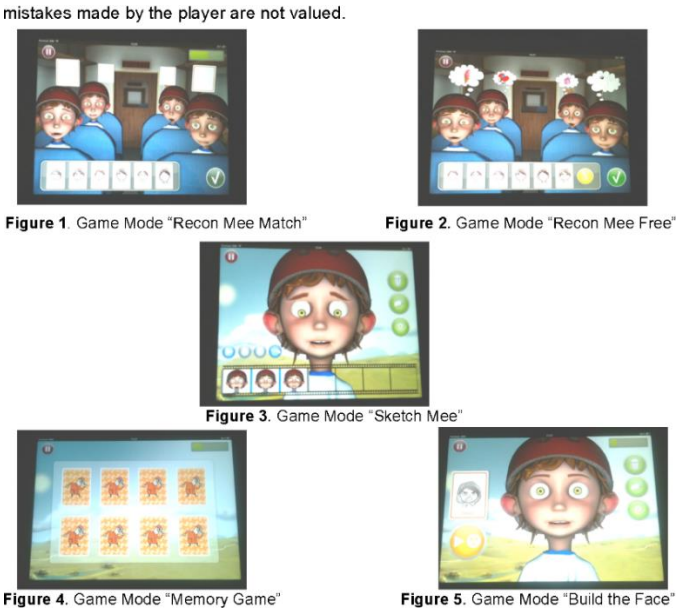
Η καινοτομία της εργασίας βρίσκεται στον συνδυασμό αυτών των εργαλείων με έναν διαδραστικό και ευχάριστο τρόπο, ο οποίος μπορεί να βοηθήσει τους θεράποντες ιατρούς να καταλάβουν καλύτερα τους ασθενείς και τα προβλήματα τους, καθώς και τους ίδιους τους ασθενείς, χωρίς το μονότονο και ,πολλές φορές, κουραστικό κομμάτι

των αυστηρών ιατρικών εξετάσεων. Παρακάτω θα παρουσιαστεί η βιβλιογραφία η οποία βοήθησε στην εύρεση των σωστών και αποδοτικών μεθόδων για την υλοποίηση της εργασίας, την δομή του νευρωνικού δικτύου και του ίδιου του παιχνιδιού, καθώς και μελλοντικές αλλαγές που μπορούν να γίνουν, για την βελτίωση της εφαρμογής.

2 *Υλικό και Μέθοδοι*

2.1 Ανάπτυξη παιχνιδιού σοβαρού σκοπού

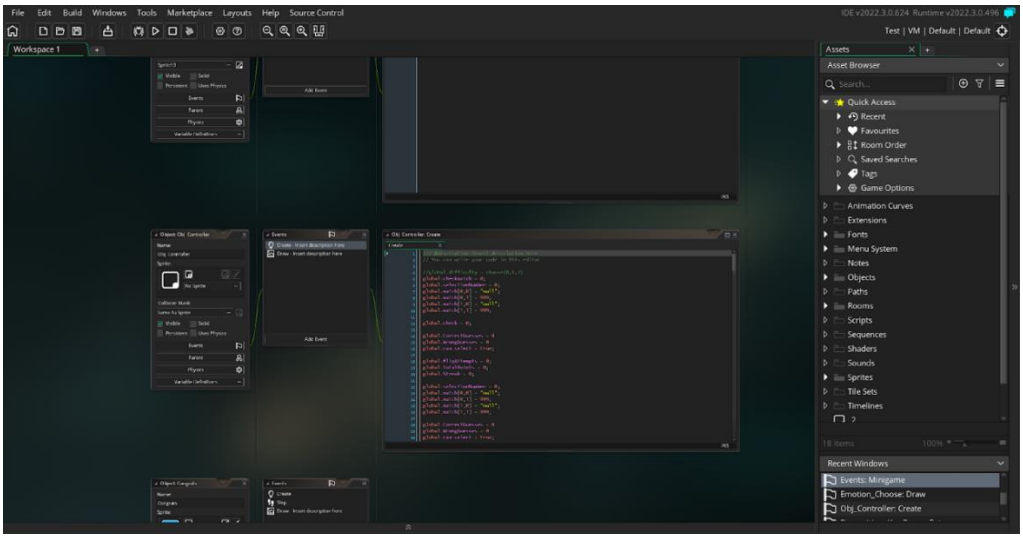
Για την αναζήτηση βιβλιογραφίας πάνω στο κομμάτι του παιχνιδιού σοβαρού σκοπού, αναζητήθηκαν άρθρα σε βάσεις δεδομένων όπως Pubmed, Google Scholar και Science Direct. Η αναζήτηση επικεντρώθηκε στην συλλογή άρθρων που αναφέρονται κυρίως σε άτομα με ψυχικές διαταραχές και σε υλοποιήσεις οι οποίες με κάποιον τρόπο περιλαμβάνουν ένα διαδραστικό παιχνίδι σοβαρού σκοπού [5]. Από την βιβλιογραφία παρατηρήθηκε ότι υπάρχει μεγάλη ποικιλία τέτοιων παιχνιδιών από παιχνίδια τύπου adventure μέχρι και παιχνίδια τύπου shooter. Η πλειοψηφία των παιχνιδιών που προορίζονται για χρήση από άτομα μικρής ηλικίας, παρόλα αυτά, φαίνεται ότι έπεφταν στην κατηγορία των puzzle παιχνιδιών [6], τα οποία ζητάνε από τον χρήστη να εκτελέσει κάποιες εύκολες ενέργειες για να λύσει ένα πρόβλημα και δεν είναι "επιβαρυντικά" ως προς τον παίκτη. Ειδικότερα, έμπνευση για το παιχνίδι της εργασίας αποτέλεσε το παιχνίδι σοβαρού σκοπού LifeIsGame, το οποίο απευθύνεται σε παιδιά με αυτισμό και σκοπεύει στο να τους ενισχύσει την συναισθηματική νοημοσύνη με διάφορα υποπαιχνίδια [7].



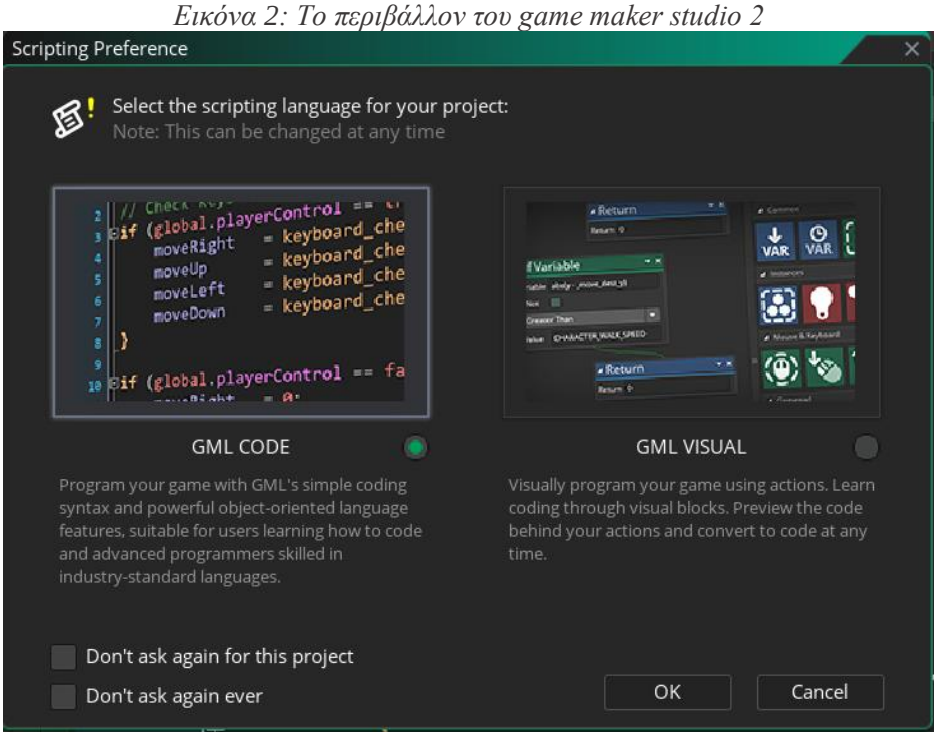
Εικόνα 1: Τα διάφορα game modes του παιχνιδιού "Life is Game" [7]

Για την υλοποίηση του παιχνιδιού χρειαζόταν να γίνει μια επιλογή ενός game engine ανάλογα με το ποιο ήταν καλύτερο για την υλοποίηση της εργασίας. Ανάμεσα σε πολλά, όπως Unity, Godot, Buildbox, έγινε η επιλογή για χρήση Game Maker και , ειδικότερα, Game Maker Studio 2, καθώς η ομάδα διέθετε περισσότερη εμπειρία με C-like γλώσσες και το συγκεκριμένο «game engine» ήταν εύκολο στην εξοικείωση και στην κατανόηση. Το Game Maker διαθέτει δύο τρόπους υλοποίησης παιχνιδιών, με GML γλώσσα, μια αυτόνομη γλώσσα του GameMaker που βασίζεται σε C# και C++ , και με GML Visualization με χρήση μπλοκ εντολών. Για την εργασία χρησιμοποιήθηκε η GML γλώσσα και η version v2022.3.0.624 του GameMaker Studio[1].

[1] www.gamemaker.io/en/gamemaker



Εικόνα 2: Το περιβάλλον του game maker studio 2



Εικόνα 3: Προγραμματιστικές επιλογές game maker 2

2.2 Ανάπτυξη Νευρωνικού Δικτύου

Για την αναζήτηση βιβλιογραφίας πάνω στο κομμάτι του νευρωνικού, αναζητήθηκαν άρθρα σε βάσεις δεδομένων όπως Google Scholar, SpringerLink, IEEE Xplore, Elsevier και Science Direct. Η αναζήτηση επικεντρώθηκε στην συλλογή άρθρων που αναφέρονται κυρίως σε ανάπτυξη νευρωνικού συστήματος για αναγνώριση φωνής και πιο ειδικά, για αναγνώριση συναισθήματος (Speech Emotion Recognition - SER). Από την βιβλιογραφία παρατηρήθηκε ότι υπάρχει μεγάλη ποικιλία τέτοιων νευρωνικών συστημάτων [1].

Σύμφωνα με την υπάρχουσα βιβλιογραφία το πιο διαδεδομένο και επιτυχημένο μοντέλο είναι το δίκτυο μακράς βραχύχρονης μνήμης (Long Short-Term Memory - LSTM), όπως αναδεικνύεται και στην εικόνα 5, το οποίο συνηθίζεται να χρησιμοποιεί σαν χαρακτηριστικό εισόδου το Mel Frequency Cepstrum Coefficients (MFCCs) [1], [2], [3], [4]. Τέλος, οι βάσεις δεδομένων που επιλέχθηκαν είναι οι εξής: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Toronto Emotional Speech Set (TESS), την Surrey Audio Expressed Emotion (SAVEE) την Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D). Παρακάτω ακολουθεί μία σύντομη ανασκόπηση για κάθε πτυχή του νευρωνικού δικτύου.

Model number		1	2	3	4	5	6	7	8	9	10	11	12
Precision	Neutral	0.89	0.76	0.7	0.83	0.7	0.87	0.79	1	0.83	0.89	0.78	0
	Angry	0.88	0.84	0.69	0.75	0.86	0.82	0.76	0.72	0.8	0.84	0.92	0.84
	Happy	0.62	0.73	0.81	0.58	0.81	0.65	0.56	0.68	0.69	0.59	0.85	0.8
	Sad	0.55	0.56	0.47	0.52	0.58	0.6	0.56	0.4	0.61	0.78	0.6	0.33
Recall	Neutral	0.61	0.68	0.5	0.36	0.57	0.71	0.54	0.14	0.36	0.54	0.5	0
	Angry	0.67	0.76	0.79	0.71	0.76	0.76	0.69	0.74	0.76	0.79	0.83	0.64
	Happy	0.71	0.69	0.74	0.6	0.71	0.69	0.71	0.54	0.83	0.89	0.8	0.57
	Sad	0.8	0.73	0.5	0.57	0.83	0.7	0.6	0.6	0.77	0.87	0.9	0.77
F1 score	Neutral	0.72	0.72	0.58	0.5	0.63	0.78	0.64	0.25	0.5	0.67	0.61	0
	Angry	0.76	0.8	0.73	0.73	0.81	0.79	0.72	0.73	0.78	0.84	0.88	0.73
	Happy	0.67	0.71	0.78	0.59	0.76	0.67	0.63	0.6	0.75	0.86	0.82	0.67
	Sad	0.65	0.64	0.48	0.54	0.68	0.65	0.58	0.48	0.68	0.7	0.72	0.46
Accuracy	Neutral	0.9	0.89	0.84	0.84	0.85	0.93	0.87	0.77	0.86	0.89	0.86	0.73
	Angry	0.87	0.88	0.82	0.84	0.89	0.88	0.84	0.83	0.87	0.9	0.93	0.85
	Happy	0.81	0.82	0.89	0.79	0.88	0.82	0.78	0.81	0.86	0.93	0.91	0.85
	Sad	0.81	0.79	0.76	0.79	0.83	0.83	0.81	0.71	0.84	0.84	0.84	0.61

Εικόνα 4: LSTM μοντέλα πάνω στην βάση RAVDESS με χρήση MFCC [8]

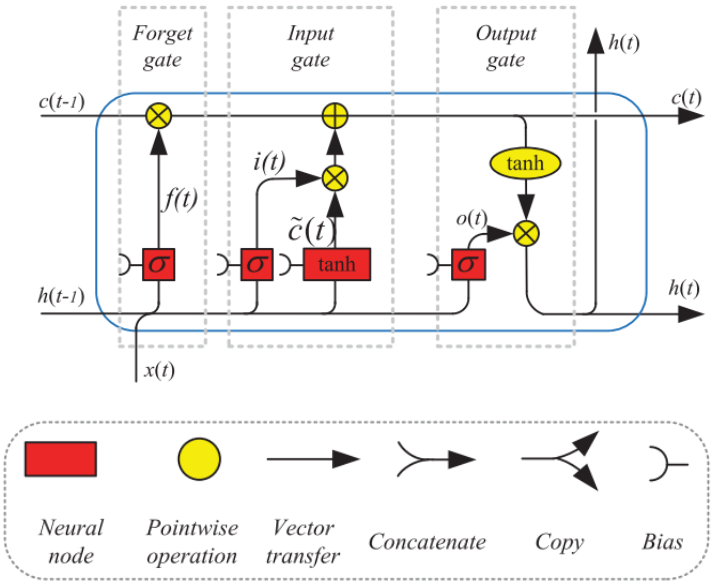
Research Title	Methodology and Number of Layers	Features	Dataset and Accuracy

Multi-Conditioning & Data Augmentation using Generative Noise Model for Speech Emotion Recognition in Noisy Conditions, Tiwari et al., 2020 [76]	• DNN/3 • Generative	• HLDs (mean, standard deviation, skewness, kurtosis, extremes, linear regressions • LLDs (zero-crossing rate (ZCR), RMS energy, F0, HNR, MFCCs)	• EMO-DB: 82.73% • IEMOCAP: 62.74%
A First Look Into A Convolutional Neural Network For Speech Emotion Detection, Bertero, and Fung, 2017 [21]	• CNN/2	• PCM	• TEDLIUM2: 66.1%
Negative Emotion Recognition using Deep Learning for Thai Language, Mekruksavanich et al., 2020 [79]	• DCNN/6	• MFCC	• SAVEE: 65.83% • RAVDESS: 75.83% • TESS: 55.71% • CREMA-D: 65.77% • THAI: 96.60%
Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching, Zhang et al., 2017 [25]	• DCNN (AlexNet)/8 • DTPM/3 • SVM	• Log Mel-Spectrogram • Delta • Delta delta	• EMO-DB: 87.31% • RML: 75.34% • eNTERFACE05: 79.25% • BAUM-1s: 44.61%
Speech Emotion Recognition with Deep Learning, Harar et al., 2017 [26]	• DCNN/10	• PCM	• EMO-DB: 96.97%
LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework, Wöllmer et al., 2013 [28]	• LSTM/1 • BLSTM/1	• Loudness, ZCR, Energy in 250–650 Hz, 1–4 kHz • 25%, 50%, 75%, and 90% spectral roll-off points, flux, entropy, variance, skewness • Psychoacoustic sharpness, harmonicity, 10 MFCCs • F0 (SHS followed by Viterbi smoothing) • Voicing, jitter, shimmer (local), delta jitter • Logarithmic Harmonics-to-Noise Ratio (logHNR)	• SEMAINE: 71.2%
Adieu Features? End-To-End Speech Emotion Recognition Using A Deep Convolutional Recurrent Network, Trigeorgis et al., 2016 [29]	• DCNN, LSTM/4	• PCM	• RECOLA: 68.4%
Speech Emotion Recognition using deep 1D & 2D CNN LSTM networks, Zhao et al., 2019 [30]	• DCNN, LSTM/5	• PCM • Log-Mel Spectrogram	• EMO-DB: 95.33% • IEMOCAP: 86.16%
Speech Emotion Classification Using Attention-Based LSTM, Xie et al., 2019 [83]	• LSTM, DNN 5	• F0, F0 envelope, ERMS noise and harmonics, and HNR • Voicing probability, ZCS, Loudness and Delta • Local jitter and shimmer, DDP jitter • MFCC and Delta, Mel spectral and logMel bands • LPC coefficients, Linear spectral pair frequency	• eNTERFACE: 89.6% • GEMEP: 57.0% • CASIA: 92.8%
Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study, Latif et al., 2018 [31]	• VAE, LSTM 2, 4	• Log-Mel Spectrogram	• IEMOCAP: 64.93%
Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition, Eskimez et al., 2018 [32]	• CNN, VAE/5, 6, 4, 10, 5	• Log-Mel Spectrogram	• IEMOCAP: 48.54%
Towards Speech Emotion Recognition "in the wild" using Aggregated Corpora and Deep Multi-Task Learning, Kim et al., 2017 [33]	• LSTM, MTL/3, 3, 2	• F0, voice probability, zero-crossing-rate • 12 MFCCs with energy and their first-time derivatives	• EMO-DB: 92.5% • eNTERFACE: 95.3% • LDC: 56.4% • Aibo: 52.0% • IEMOCAP: 56.9%
Adversarial Machine Learning and Speech Emotion Recognition: Utilizing Generative Adversarial Networks for Robustness, Latif et al., 2018 [86]	• LSTM, GAN/2	• eGeMAPS features	• Aibo: 64.86% • IEMOCAP: 53.76%
On Enhancing Speech Emotion Recognition Using Generative Adversarial Networks, Sahu et al., 2018 [88]	• GAN, SVM	• 1582-dimensional openSMILE feature space	• IEMOCAP: 60.29%
Data Augmentation Using GANs for Speech Emotion Recognition, Chatziagapi et al., 2019 [89]	• DCNN (VGG19), GAN/19	• 128 MFCCs	• IEMOCAP: 53.6% • Feel-25k: 54.6%
human–computer Using Transfer Learning, Song et al., 2014 [90]	• PCA, LPP, TSL	• 12 MFCCs and Delta • 8 LSF, Intense, Loudness, ZCR • Voice probability, F0, F0 envelopes	• EMO-DB: 59.8%
Transfer Linear Subspace Learning for Cross-Corpus Speech Emotion Recognition, Song 2019 [91]	• LDA, TSL, TLSL	• 1582-dimensional openSMILE feature space	• EMO-DB, eNTERFACE, Aibo: 54.61%
Automatic Speech Emotion Recognition using recurrent neural networks local attention, Mirsamadi et al., 2017 [94]	• LSTM, ATTN/4, 3, 3, 4, 4, 4	• 257-dimensional magnitude FFT vectors • F0, voice probability, frame energy, ZCR • 12 MFCCs and Delta	• IEMOCAP: 63.5%

Εικόνα 5: Μια σύντομη σύγκριση κατά μέθοδο που χρησιμοποιήθηκε (εάν υπάρχουν, τον αριθμό των επιπέδων/στρωμάτων-layers), χαρακτηριστικά που χρησιμοποιήθηκαν και υψηλότερη ακρίβεια που αναφέρεται για κάθε σύνολο δεδομένων όλων των αλγορίθμων που αναφέρονται στο [1]

2.2.1 LSTM Μοντέλο

Το LSTM είναι ένα RNN δίκτυο που δημιουργήθηκε από τους Hochreiter και Schmidhuber το 1997 [9]. Σκοπός του είναι να επιλύσει τα vanishing gradient προβλήματα που παρατηρούνται σε άλλες διατάξεις RNNs- μέσω προσθήκης μηχανισμών για τη ρύθμιση των πληροφοριών που επιτρέπουν τη διατήρησή τους για μεγάλες χρονικές περιόδους [10]. Συγκεκριμένα, το LSTM μοντέλο περιλαμβάνει πύλες για τον έλεγχο της ροής πληροφοριών μεταξύ των κελιών. Οι πληροφορίες που ρέουν κατά μήκος της εκάστοτε κατάστασης ενός κελιού μπορούν να διαμορφωθούν από τις ενδιάμεσες πύλες τύπου input και forget. Η τελική έξοδος είναι μια φιλτραρισμένη έκδοση της κατάστασης κελιού με βάση το περιεχόμενο των τιμών εισόδου.

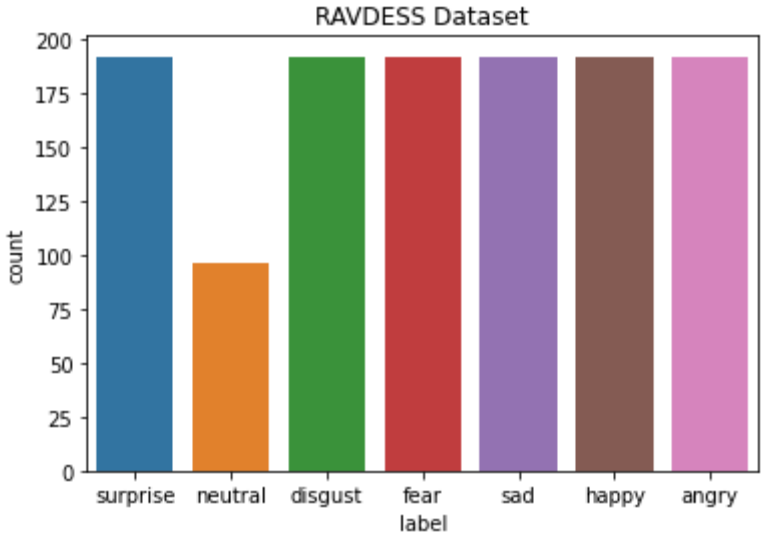


Εικόνα 6: Αρχιτεκτονική LSTM cell [11]

2.2.2 Βάσεις Δεδομένων

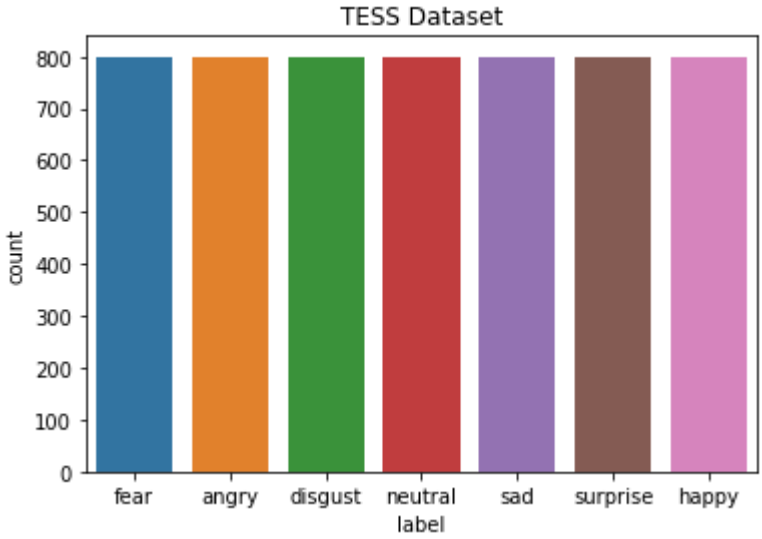
Το βασικό κριτήριο επιλογής των βάσεων δεδομένων που χρησιμοποιήθηκαν στην εκπαίδευση και δοκιμή του μοντέλου, ήταν η αναφορά και η χρήση τους σε πολλές έρευνες, καθιστώντας τες αξιόπιστες για χρήση. Επιπλέον, τέθηκε ως κριτήριο η ύπαρξη μίας κοινής γλώσσας αναφοράς-της αγγλικής-μιας και έχει παρατηρηθεί ότι συναισθήματα όπως η «αηδία» και ο «φόβος» εξαρτώνται αρκετά από το πολιτισμικό, υπόβαθρο, καθώς και από περιβαλλοντικές ή κοινοτικές διαφορές [12]. Έτσι, η προσθήκη ακόμα ενός αστάθμητου παράγοντα κατά την ανάπτυξη του μοντέλου θα καθιστούσε δυσκολότερη την ανάλυση της εγκυρότητας των αποτελεσμάτων.

Η βάση δεδομένων RAVDESS [13] είναι μια επικυρωμένη πολυμορφική βάση δεδομένων συναισθηματικού λόγου και τραγουδιού, της οποίας τα δεδομένα είναι ισορροπημένα μεταξύ των 2 φύλων και περιλαμβάνει 24 επαγγελματίες ηθοποιούς. Η ομιλία περιλαμβάνει τα εξής οκτώ labels φωνής: ήρεμη, χαρούμενη, λυπημένη, θυμωμένη, εκφράσεις φόβου, έκπληξης και αηδίας και το τραγούδι περιέχει ήρεμα, χαρούμενα, λυπημένα, θυμωμένα και φοβισμένα συναισθήματα. Κάθε έκφραση παράγεται σε δύο επίπεδα συναισθηματικής έντασης, με μια επιπλέον ουδέτερη έκφραση. Παρά το μεγάλο πλήθος δεδομένων και την ισορροπία τους, η βάση RAVDESS οδηγεί σε σχετικά χαμηλά ποσοστά ακρίβειας, εξαιτίας του γεγονότος ότι ο αριθμός των δειγμάτων στην ουδέτερη κατηγορία είναι ο μισός από αυτόν που βρίσκεται σε άλλες κλάσεις με αποτέλεσμα το μοντέλο να αδυνατεί να εκπαιδευτεί στην αναγνώριση των χαρακτηριστικών της ουδέτερης τάξης [14].



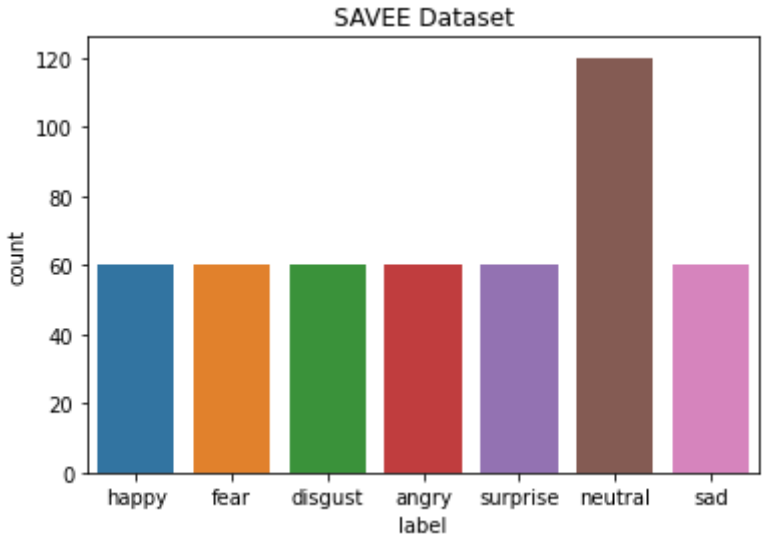
Εικόνα 7: Βάση Δεδομένων RAVDESS

Η βάση δεδομένων TESS [15] περιλαμβάνει συνολικά 2800 ερεθίσματα, στα οποία περιέχεται ένα σύνολο 200 λέξεων-στόχων από δύο γυναίκες ηθοποιούς από την περιοχή του Toronto (ηλικίας 26 και 64 ετών), ενώ έγιναν ηχογραφήσεις του σετ που απεικονίζει καθένα από τα επτά συναισθήματα (θυμός, αηδία, φόβος, ευτυχία, ευχάριστο έκπληξη, θλίψη και ουδέτερο). Σημειώνεται ότι η απουσία δειγμάτων από ανδρικές φωνές αποτελεί έναν παράγοντα ο οποίος θα μπορούσε να οδηγήσει σε χαμηλά στατιστικά scores.



Εικόνα 8: Βάση Δεδομένων TESS

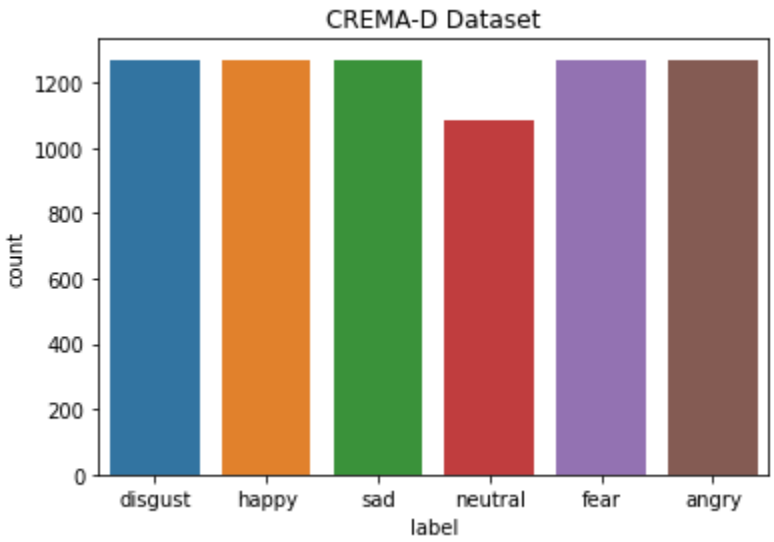
Η βάση δεδομένων SAVEE [16] περιλαμβάνει 480 εκφράσεις από τέσσερις αγγλόφωνους άνδρες, οι οποίες κατηγοριοποιούνται σε επτά συναισθήματα, ενώ οι προτάσεις επιλέχθηκαν από το σώμα TIMIT και ήταν φωνητικά ισορροπημένες για κάθε συναίσθημα. Αξίζει να σημειωθεί ότι η SAVEE βιβλιοθήκη επιλέχθηκε ώστε να αντισταθμίσει την βάση δεδομένων TESS, η οποία περιλαμβάνει μόνο δείγματα φωνής από γυναίκες ηθοποιούς.



Εικόνα 9: Βάση Δεδομένων SAVEE

Τέλος, η βάση δεδομένων CREMA-D [17] αποτελείται από 7.442 δείγματα εκφράσεων προσώπου και φωνητικών συναισθημάτων σε προτάσεις που εκφωνούνται από 91 ηθοποιούς διαφορετικών εθνοτήτων σε μια σειρά βασικών συναισθηματικών καταστάσεων (χαρούμενος, λυπημένος, θυμός, φόβος, αηδία και ουδέτερη). Κάθε

δείγμα σε αυτό το σύνολο δεδομένων έχει δύο αξιολογήσεις, μία για την κατηγορία των συναισθημάτων και τη δεύτερη για την έντασή του. Η επισήμανση σε αυτό το σύνολο δεδομένων έγινε με τη συγκέντρωση 223.260 μεμονωμένων αξιολογήσεων μέσω 2443 μεμονωμένων αξιολογητών. Το μεγάλο πλήθος ηθοποιών και δειγμάτων, σε συνδυασμό με την ικανοποιητική ισορροπία μεταξύ των ετικετών καθιστά την εν λόγω βιβλιοθήκη ως μία από τις καταλληλότερες για την εκμάθηση του μοντέλου.

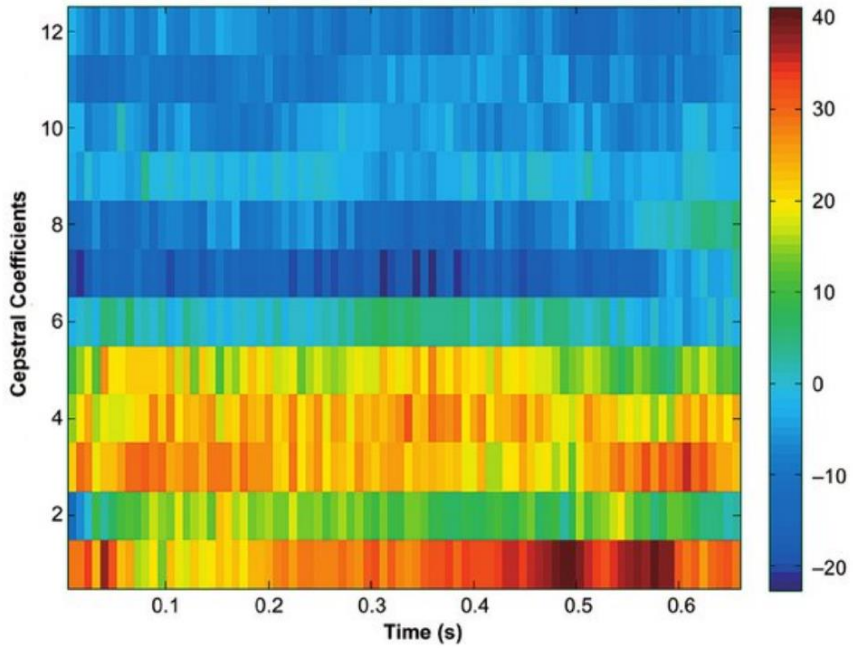


Εικόνα 10: Βάση Δεδομένων CREMA-D

2.2.3 Mel Frequency Cepstrum Coefficients

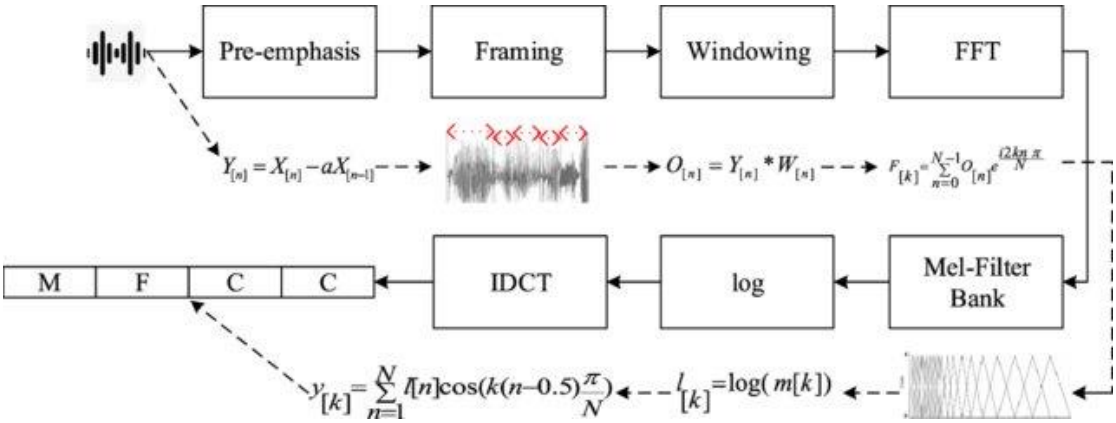
Ως είσοδοι του νευρωνικού δικτύου, χρησιμοποιήθηκαν τα Mel Frequency Cepstrum Coefficients (MFCC), που χρησιμοποιούνται για εργασίες που σχετίζονται με την ομιλία, όπως η αναγνώριση φύλου με φωνή, η αναγνώριση ομιλίας, η αναγνώριση συναισθημάτων ομιλίας και πολλά άλλα [18]. Οι συντελεστές MFCC είναι επιτυχείς επειδή προέρχονται από τα πρότυπα ανθρώπινης ομιλίας [19]. Το MFCC επιχειρεί να μιμηθεί το ανθρώπινο αυτί, όπου η ακουστική συχνότητα προσδιορίζεται ως ασύμμετρο φάσμα [20].

Η κλίμακα του Mel είναι η κλίμακα αντίληψης των pitches που το κοινό θεωρεί ίσο με την απόσταση μεταξύ τους. Το Mel προέρχεται από τη λέξη μελωδία(melody), που δείχνει ότι η κλίμακα βασίζεται σε συγκρίσεις τόνου. Το σημείο αναφοράς μεταξύ αυτής της κλίμακας και της μέτρησης της κανονικής συχνότητας ορίζεται ως ένας τόνος 1000 Hz, ίσος με 40 dB πάνω από το κατώφλι του ακροατή, με ύψος 1000 mels.



Εικόνα 11: Mel-frequency cepstral coefficients (MFCC)—θορυβοποιημένη αντρική φωνή [21]

Ο αλγόριθμος [22] από τον οποίον εξάγονται τα coefficients περιλαμβάνει παραθυροποίηση του σήματος εισόδου, εφαρμογή Discrete Fourier Transform(DFT), προσαρμογή των συχνοτικών δειγμάτων σε Mel scale και τέλος εφαρμογή inverse Discrete Cosine Tranform (DCT), όπως φαίνεται και στο κάτωθι διάγραμμα



Εικόνα 12: Βασικά βήματα εφαρμογής του MFCC αλγορίθμου [23]

2.2.4 Προγραμματιστική Υλοποίηση

Τέλος, αναφορικά με την προγραμματιστική υλοποίηση του νευρωνικού, χρησιμοποιήθηκαν η γλώσσα προγραμματισμού python, καθώς και οι βιβλιοθήκες keras [24] και sklearn [25], πάνω στις οποίες υπάρχει πληθώρα εφαρμογών και βιβλιογραφίας για την μοντελοποίηση LSTM μοντέλων [26].

3 Αποτελέσματα

3.1 Ανάπτυξη παιχνιδιού σοβαρού σκοπού

Το παιχνίδι με τίτλο “The Cardomancer’s Lair” είναι ένα παιχνίδι καρτών τύπου Concentration[2] στο οποίο ο παίκτης καλείται να ταιριάξει ίδιες κάρτες και αμέσως μετά από ένα σωστό ταίριασμα να ερμηνεύσει το συναίσθημα που παρουσιάζεται πάνω στις εικόνες. Παράλληλα, δεδομένης της εισόδου του νευρωνικού δικτύου, το παιχνίδι έχει σκοπό να αλληλεπιδρά ανάλογα με το συναίσθημα του χρήστη μέσα από κάποια συμβάντα, ως τρόπος ενίσχυσης της συναισθηματικής νοημοσύνης και εκμάθησης των συναισθημάτων.

Εικόνα 13: Το αρχικό μενού του παιχνιδιού

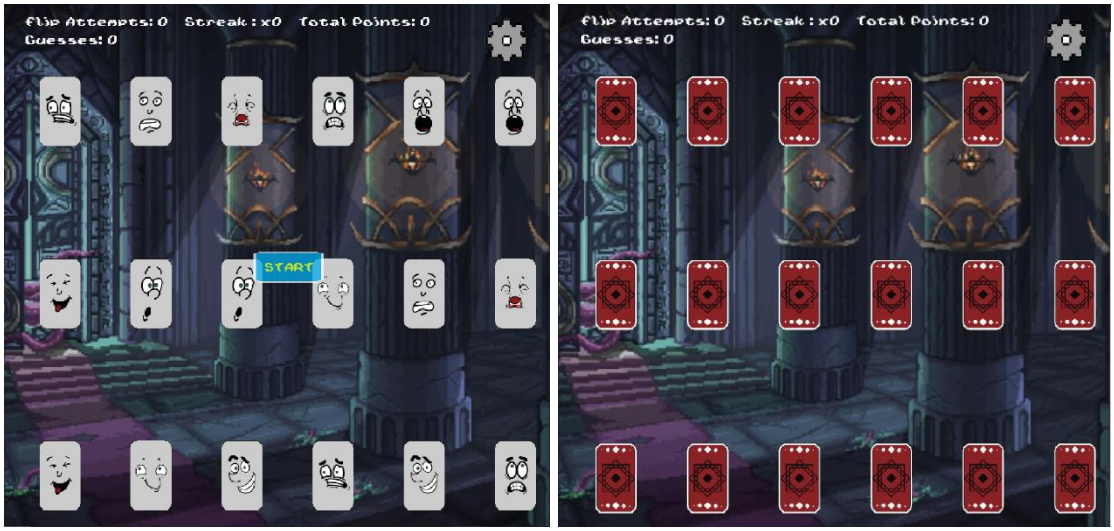
Εικόνα 14: Μενού ρυθμίσεων

Στο αρχικό πλαίσιο του παιχνιδιού δίνεται στον παίκτη η επιλογή να ξεκινήσει μία συνεδρία του παιχνιδιού, η επιλογή των “Ρυθμίσεων” και η επιλογή να κλείσει την εφαρμογή. Στην “Ρυθμίσεις” του παιχνιδιού ο παίκτης διαθέτει την ικανότητα να αλλάξει την δυσκολία του παιχνιδιού, οι οποίες διαφοροποιούν τον αριθμό των καρτών

[2] https://en.wikipedia.org/wiki/Concentration_(card_game)

της συνεδρίας, και παρέχεται η επιλογή εισόδου του συναισθήματος από τον υπεύθυνο παρακολούθησης.

Στην εικόνα 15 φαίνεται το περιβάλλον του παιχνιδιού με παράδειγμα συνεδρίας σε "Μεσαία" δυσκολία.



Εικόνα 15: Περιβάλλον του παιχνιδιού με παράδειγμα συνεδρίας σε "Μεσαία" δυσκολία

Η κρυφή μεριά των καρτών παρουσιάζεται αρχικά στο χρήστη και ύστερα από την επιλογή του κουμπιού "Εκκίνησης" οι κάρτες γυρνάνε ανάποδα και ξεκινάει η συνεδρία.

Καθ' όλη την διάρκεια του παιχνιδιού παρέχονται οι εξής πληροφορίες στον χρήστη:

1. Οι προσπάθειες που έχει κάνει για την εύρεση ταιριάσματος
2. Το σερί που έχει εκτελέσει στην εύρεση των καρτών
3. Οι πόντοι που έχει μαζέψει
4. Οι σωστές μαντεψιές

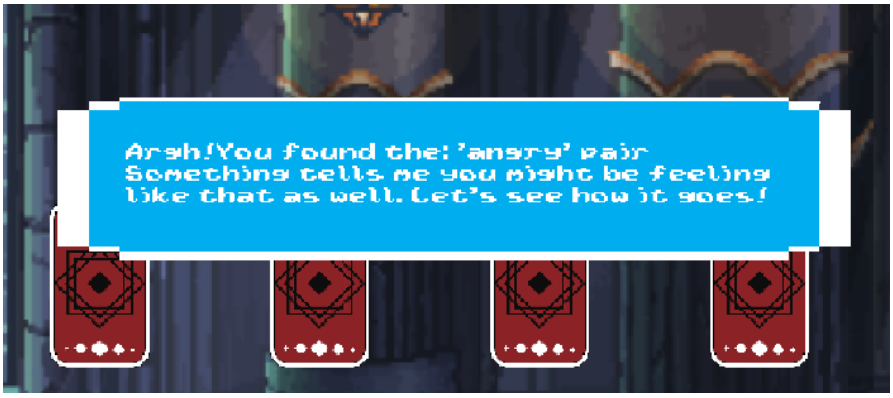
Το "Γρανάζι" λειτουργεί ως κουμπί ρυθμίσεων και παρέχει στον χρήστη την δυνατότητα να επιστρέψει στο αρχικό μενού, να κλείσει το παιχνίδι ή να κάνει επανεκκίνηση της συνεδρίας.

Εικόνα 16: Επιλογές κουμπιού ρυθμίσεων

Κατά την διάρκεια, λοιπόν, της συνεδρίας , ο παίκτης ταιριάζει μεταξύ τους κάρτες και σε περίπτωση σωστού ταιριάσματος του δίνεται η επιλογή να επιλέξει το συναίσθημα που παρουσιάζεται στην κάρτα.

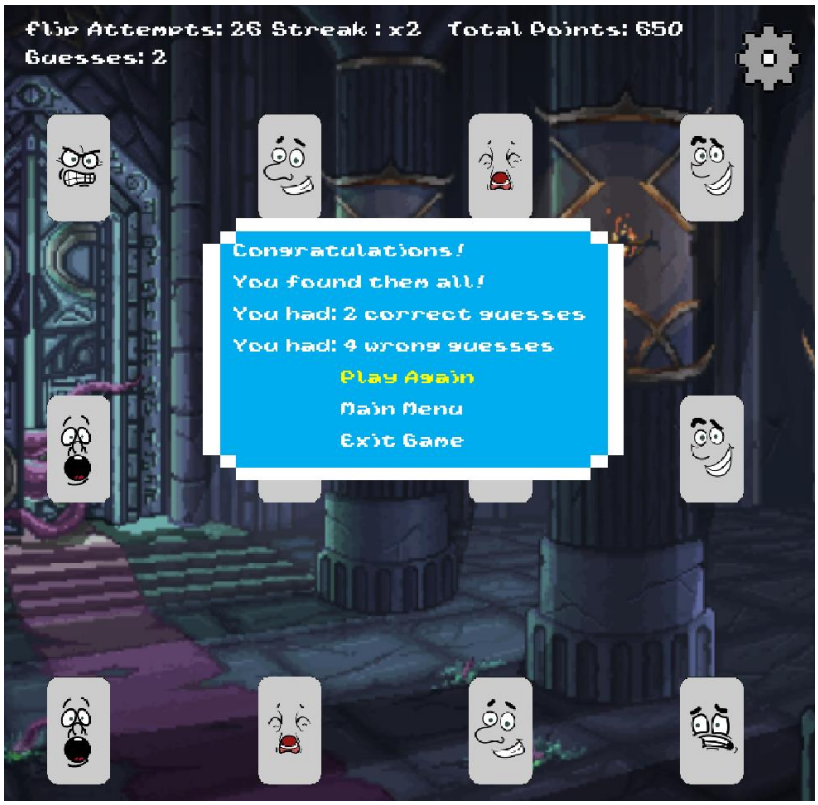
Εικόνα 17: Ο παίκτης καλείται να επιλέξει ποιο συναίσθημα εκφράζουν οι κάρτες

Στην περίπτωση της εικόνας 17, η σωστή επιλογή είναι η τρίτη, ο “Θυμός”. Αν το συναίσθημα που έχει επιλεχθεί από το αρχικό μενού συμπίπτει με αυτό που εκφράζει το ζευγάρι των καρτών τότε εμφανίζεται το μήνυμα που φαίνεται στην εικόνα 18.



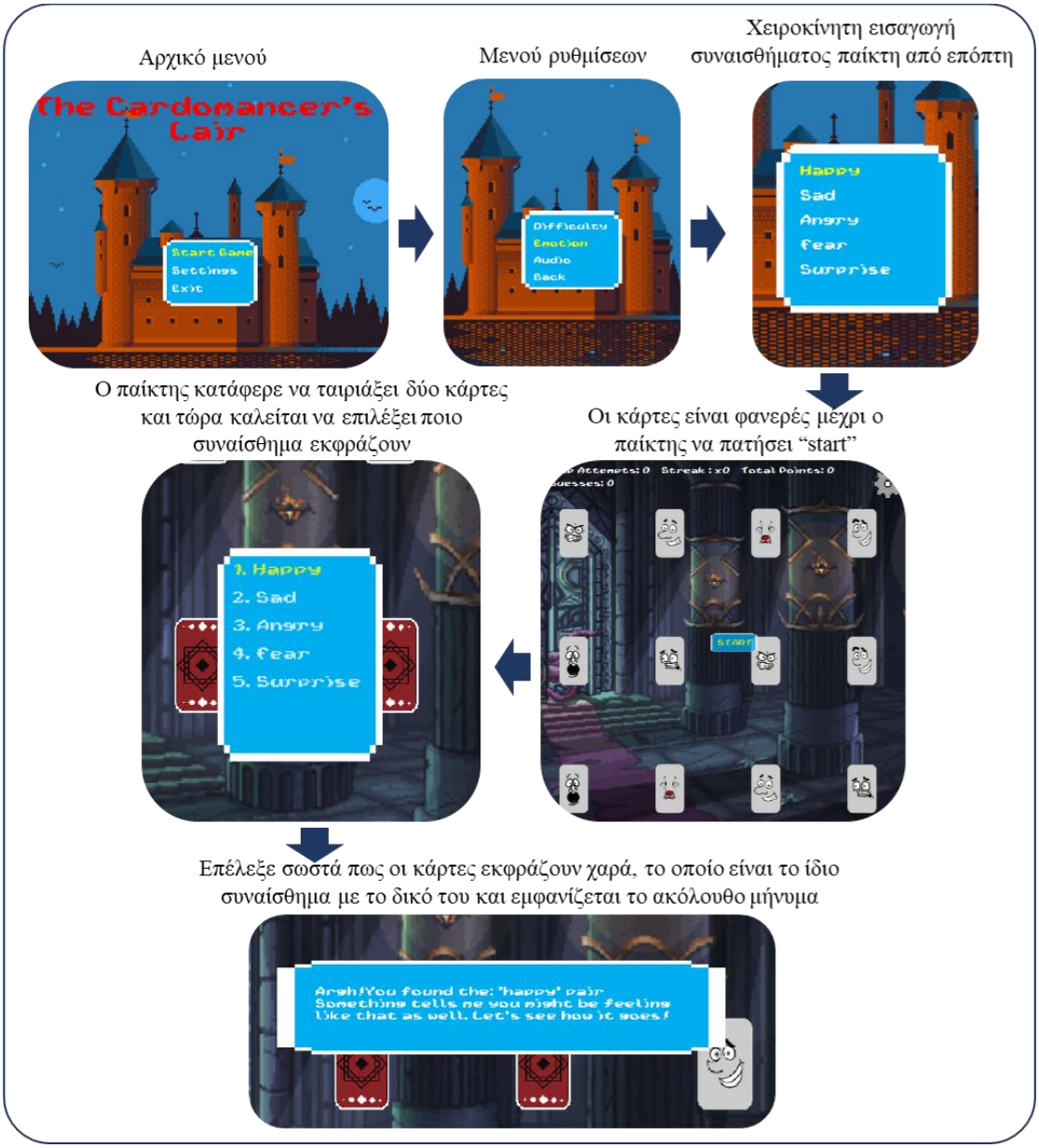
Εικόνα 18: Το μήνυμα που εμφανίζεται σε περίπτωση ταιριάσματος κάρτας που εκφράζουν το συναίσθημα του παίκτη

Ο κύριος σκοπός είναι να αναγνωρίζει το παιχνίδι τα συναισθήματα του χρήστη και με τις αλληλεπιδράσεις να μαθαίνει και στον ίδιο με παράδειγμα τον εαυτό του το πώς αισθάνεται. Με αυτόν τον τρόπο, ενισχύεται η συναισθηματική του νοημοσύνη με γνώμονα τον ίδιο του τον εαυτό και αυτό που αισθάνεται κατά την διάρκεια της συνεδρίας. Στο τέλος της συνεδρίας, παρουσιάζεται το παράθυρο που φαίνεται στην εικόνα 19.



Εικόνα 19: Τέλος συνεδρίας

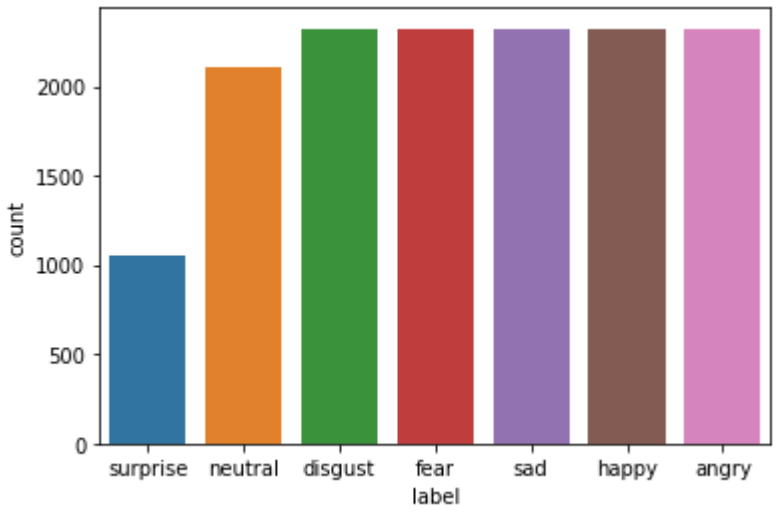
Το παράθυρο παρέχει πληροφορίες οι οποίες μπορεί να είναι χρήσιμες στον θεράποντα ιατρό. Οι "Σωστές Μαντεψιές" υποδηλώνουν το πόσο καλά μπορεί να καταλάβει τα συναισθήματα, ενώ οι "Λάθος Μαντεψιές" υποδεικνύουν τον προβληματισμό του ασθενή σε αυτήν την αναγνώριση. Κατά την διάρκεια του παιχνιδιού μπορεί να διατηρείται επίσης, η πληροφορία για το ποια συναισθήματα δεν αναγνώρισε σωστά, γεγονός που μπορεί να είναι χρήσιμο στην περαιτέρω εκμάθηση του.



Εικόνα 20: Συνολική ροή μίας συνεδρίας του παιχνιδιού

3.2 Ανάπτυξη Νευρωνικού Δικτύου

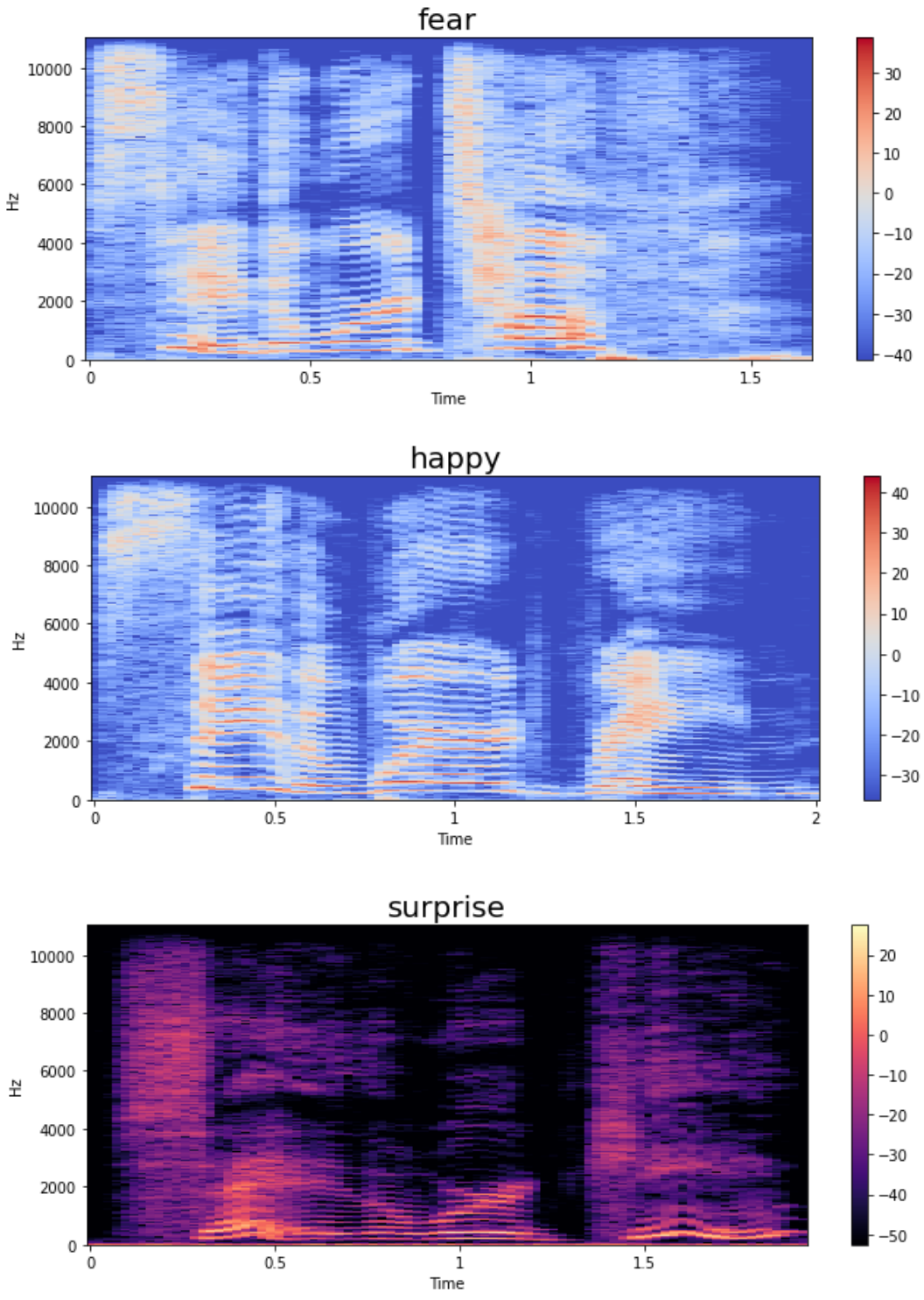
Το νευρωνικό δίκτυο LSTM που μοντελοποιήθηκε χρησιμοποίησε για την προπόνησή του και τις 4 βάσεις δεδομένων CREMA-D, RAVDESS, TESS και SAVEE. Έτσι, συνολικά προέκυψε μία ενοποιημένη βάση με 14770 στοιχεία.



Εικόνα 21: Συνολικά δεδομένα

Αξίζει να σημειωθεί ότι για τα συναισθήματα «surprise» και «neutral» υπήρχε χαμηλότερη διαθεσιμότητα δειγμάτων από τις επιλεγμένες βάσεις, γεγονός που αναμένεται να οδηγήσει σε χαμηλότερα scores για τις εν λόγω ετικέτες.

Όσον αφορά τα δείγματα καθ' αυτά, έγινε η εξαγωγή των 40 πρώτων MFCC's από κάθε δείγμα το οποίο στην συνέχεια αποτέλεσε είσοδο του νευρωνικού δικτύου. Η εν λόγω συχνοτική ανάλυση των δειγμάτων μπορεί να καταστεί ακόμα πιο έγκυρη αν γίνουν αντιληπτές οι διαφορές μεταξύ των διαφορετικών spectrograms ανάμεσα σε δείγματα από διαφορετικές ετικέτες.



Εικόνα 22: Spectrograms ηχητικών δειγμάτων από διαφορετικά labels του dataset

Γίνεται αντιληπτό ότι τόσο το αίσθημα της χαράς όσο και το αίσθημα του φόβου έχουν αισθητά πιο υψίσυχνο περιεχόμενο από αυτό της έκπληξης, του οποίου οι τιμές είναι πιο ομοιόμορφα κατανεμημένες στον άξονα συχνοτήτων.

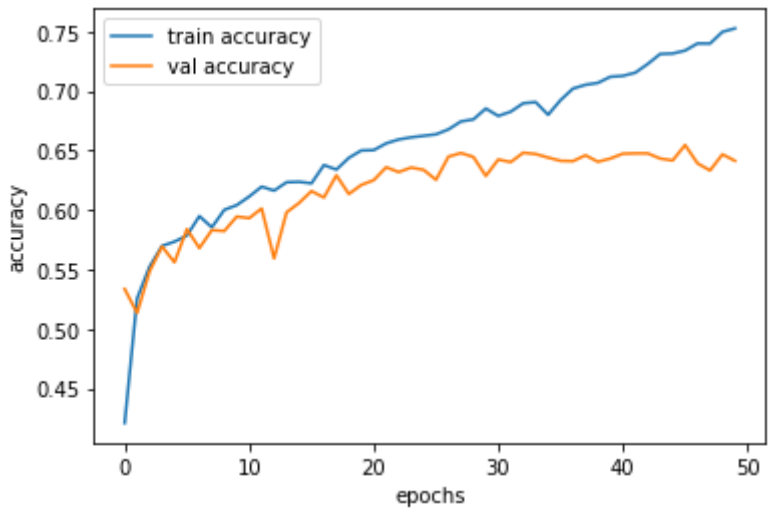
Όσον αφορά το LSTM μοντέλο καθ' αυτό, χρησιμοποιήθηκαν 3 layers με Dropout Rate 0.2, έτσι ώστε να αποφευχθούν φαινόμενα overfitting του δικτύου. Ως συναρτήσεις

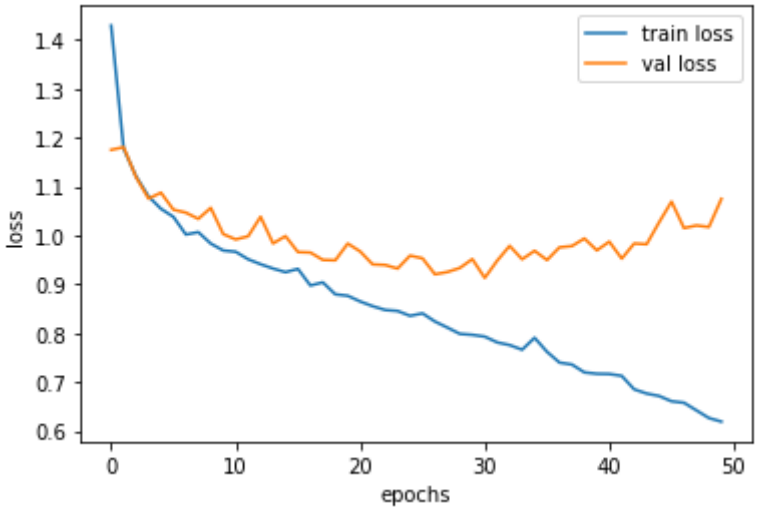
μεταφοράς για τα ενδιάμεσα layers χρησιμοποιήθηκε η ReLU Function, ενώ για το layer εξόδου η softmax, η οποία είναι καταλληλότερη για multi-clustering προβλήματα, δηλαδή προβλήματα τα οποία έχουν πάνω από 2 ετικέτες. Τέλος, το αρχικό dataset διαχωρίστηκε σε 80% training και 20% test set, ενώ η προπόνηση διήρκησε 50 εποχές(epochs).

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 256)               264192
_________________________________________________________________
dropout (Dropout)            (None, 256)               0
_________________________________________________________________
dense (Dense)                (None, 128)               32896
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_2 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 7)                 455
=================================================================
Total params: 305,799
Trainable params: 305,799
Non-trainable params: 0
```

Εικόνα 23: Χαρακτηριστικά Δομής Νευρωνικού Δικτύου

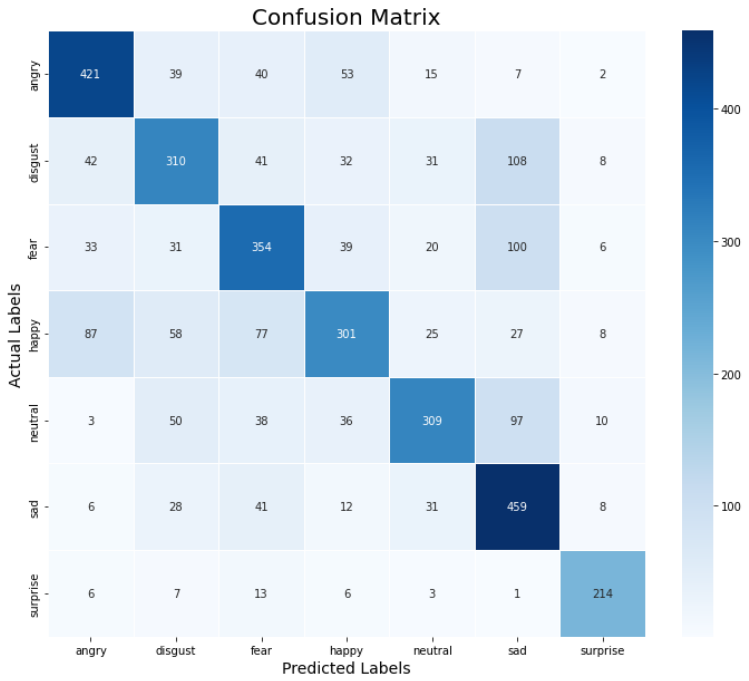
Κατά την έναρξη της εκπαίδευσης το εξεταζόμενο μοντέλο παρουσίασε χαμηλά ποσοστά ακρίβειας (42%), τα οποία όμως ανέβηκαν στο 75.2% μετά από 50 εποχές. Αντίστοιχα κατά τη διάρκεια της εκπαίδευσης παρατηρήθηκε ανάλογη πτωτική τάση στο cross-entropy loss. Το γεγονός αυτό αναδεικνύει την εγκυρότητα των παραμέτρων που εφαρμόστηκαν για την ανάπτυξη του μοντέλου.





Εικόνα 24: Spectrograms ηχητικών δειγμάτων από διαφορετικά labels του dataset

Στην συνέχεια, πραγματοποιήθηκε απεικόνιση του confusion matrix καθώς και των f1-scores ανά συναίσθημα, όπου παρατηρήθηκε αισθητά αυξημένη διακριτική ικανότητα του δικτύου για συναισθήματα όπως θυμός και έκπληξη. Αντίθετα, χαμηλά f1-scores υπήρξαν σε πιο θεμελιώδη συναισθήματα όπως ο «φόβος» και η «χαρά», γεγονός που αναδεικνύει την ανάγκη για ενίσχυση των samples στις εν λόγω ετικέτες ή και πιθανόν την ανάγκη προσθήκης περισσότερων εποχών ή περισσότερων layers επεξεργασίας της πληροφορίας στο νευρωνικό δίκτυο.



Εικόνα 25: Confusion Matrix

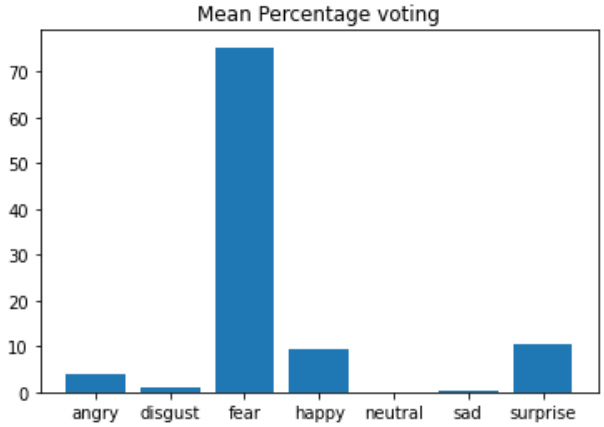
	precision	recall	f1-score	support
angry	0.70	0.73	0.72	577
disgust	0.59	0.54	0.57	572
fear	0.59	0.61	0.60	583
happy	0.63	0.52	0.57	583
neutral	0.71	0.57	0.63	543
sad	0.57	0.78	0.66	585
surprise	0.84	0.86	0.85	250
accuracy			0.64	3693
macro avg	0.66	0.66	0.66	3693
weighted avg	0.65	0.64	0.64	3693

Εικόνα 26: Classification report

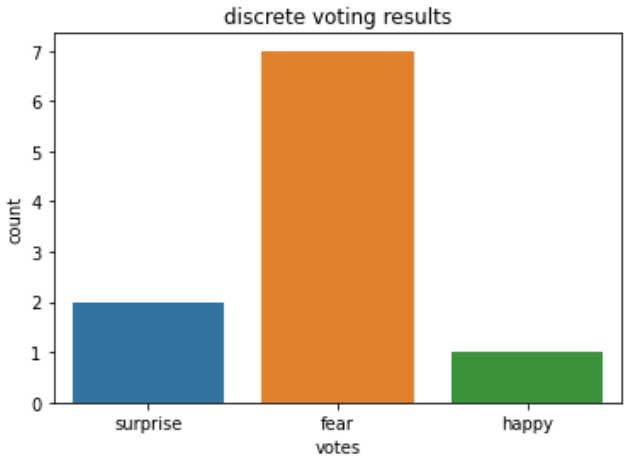
Τέλος, υλοποιήθηκε και ένα voting session, στην προσπάθεια ρεαλιστικής προσομοίωσης χρήσης του νευρωνικού στα πλαίσια του serious game. Συγκεκριμένα, έγινε λήψη δέκα διαφορετικών τυχαίων δειγμάτων από ένα τυχαίο συναίσθημα, τα οποία αντιστοιχούν σε 10 διαφορετικά δείγματα ήχου από τον παίκτη κατά την διάρκεια του παιχνιδιού. Στην συνέχεια κάθε ένα από αυτά εισάγεται στον LSTM για labeling.

Το μοντέλο έχει προσαρμοστεί έτσι ώστε να κρατάει τις 10 διαφορετικές ψήφους του νευρωνικού για κάθε ένα δείγμα (discrete screening) και έναν μέσο όρο των ποσοστών από τα αποτελέσματα του νευρωνικού για όλα τα δείγματα, από τα οποία στην συνέχεια διατηρεί την ετικέτα με το υψηλότερο ποσοστό (analog screening).

	mean %	emotions
0	3.947445	angry
1	0.947962	disgust
2	75.339783	fear
3	9.23105	happy
4	0.012954	neutral
5	0.114202	sad
6	10.406601	surprise



Εικόνα 27: Analog Voting Session για 10 δείγματα φόβου



Εικόνα 28: Discrete Voting Session για 10 δείγματα φόβου

Και στις δύο μεθόδους ψηφοφορίας για δείγματα «φόβου» το δίκτυο αποφασίζει με σχετική ευκολία την ορθή απόφαση, παρά το γεγονός ότι σε ορισμένα δείγματα αναγνωρίζει πιο πολύ άλλα συναισθήματα (εν προκειμένω έκπληξης και χαράς).

4 *Συμπεράσματα -*

Επίλογος

Η άνωθι εργασία παρουσιάζει την δυνατότητα δημιουργίας σύγχρονων εργαλείων και εφαρμογών τα οποία λειτουργούν με ευχάριστους και διαδραστικούς τρόπους για τον χρήστη, ενώ ταυτόχρονα παρέχουν απαραίτητα δεδομένα στους ειδικούς για την διάγνωση και ενίσχυση ασθενών με προβλήματα υγείας. Η ενσωμάτωση τέτοιων μεθόδων μπορεί να οδηγήσει στην δημιουργία ενός υγιούς και ήρεμο περιβάλλοντος, στο πλαίσιο του οποίου θα γίνεται η συλλογή δεδομένων, καθώς ο ασθενής δεν πιέζεται από κουραστικές ιατρικές εξετάσεις. Αξίζει να σημειωθεί, όμως, ότι αυτά τα εργαλεία δεν δύναται να αντικαταστήσουν πλήρως τις θεμελιωμένες ιατρικές εξετάσεις, αλλά θα πρέπει να διαθέτουν υποβοηθητικό ρόλο για τους ασθενείς και τους θεράποντες ιατρούς.

Παράλληλα, μία τέτοια υλοποίηση ακολουθεί πιστά το πνεύμα των σύγχρονων διαγνωστικών και θεραπευτικών μεθόδων, οι οποίες στοχεύουν στην υλοποίηση δυναμικών εξατομικευμένων συστημάτων μέσω διαρκούς προσαρμογής τους στα νέα δεδομένα που λαμβάνονται από τον ασθενή. Η εξατομίκευση αυτή στις μεθόδους διάγνωσης και θεραπείας είναι ένας σημαντικός στόχος στην ιατρική, ώστε οι διαδικασίες να μην βασίζονται σε γενικευμένες προσεγγίσεις που μπορεί να είναι βλαβερές για ορισμένους ασθενείς ή να μην μεριμνούν για ορισμένα χαρακτηριστικά. Τα νευρωνικά δίκτυα είναι εργαλεία τα οποία μπορούν να βοηθήσουν σε αυτήν την διαδικασία, καθώς είναι συστήματα τα οποία ανάλογα την είσοδο μπορούν να κατατάξουν τον ασθενή στις σωστές κατηγορίες. Όμως, τα ιατρικά δεδομένα, πάνω στα οποία βασίζονται τα νευρωνικά δίκτυα, παρουσιάζουν μεγάλα προβλήματα στην χρήση τους. Για αυτόν ακριβώς τον λόγο είναι απαραίτητο να γίνουν ραγδαίες αλλαγές στην διαχείριση των ιατρικών δεδομένων, ώστε να μπορούν τα εργαλεία αυτά να εκπαιδευτούν με ποικίλες περιπτώσεις ασθενών για να είναι πιο αποδοτικά και έμπιστα και για να αποφευχθούν περιστατικά overfitting ή underfitting.

Αναφορικά με την υλοποίησή μας, κρίνεται αναγκαία η εφαρμογή των κατάλληλων τροποποιήσεων, οι οποίες θα αφορούν τόσο την βελτιστοποίηση του νευρωνικού δικτύου αναγνώρισης της συναισθηματικής κατάστασης του χρήστη όσο και την ευρύτερη εμπειρία του χρήστη με το παιχνίδι. Πιο αναλυτικά, εξετάζουμε τις εξής βελτιώσεις:

1. Άμεση σύνδεση συστήματος αναγνώρισης συναισθημάτων με το παιχνίδι,
2. Αύξηση ακρίβειας (accuracy) και άλλων μετρικών χρησιμοποιώντας πιο αναβαθμισμένες τεχνικές επεξεργασίας δεδομένων,
3. Δοκιμή παραπάνω μοντέλων στην υλοποίηση του νευρωνικού δικτύου και εισαγωγή δεδομένων με ετικέτες βάσει των συνεδριών του παιχνιδιού,
4. Αύξηση της ποικιλίας της αλληλεπίδρασης του παίκτη με το σύστημα αναγνώρισης π.χ. λήψη ηχητικού δείγματος κατά τη διάρκεια της συνεδρίας,
5. Εισαγωγή συστημάτων για την καταπολέμηση αρνητικών συναισθημάτων και την ενίσχυση θετικών συναισθημάτων,
6. Εξαγωγή στατιστικών μετρικών για το εάν η αποτυχία στην αναγνώριση οφείλεται και σε εξωτερικούς παράγοντες, όπως στην ύπαρξη απότομων συναισθηματικών αλλαγών του χρήστη
7. Εξαγωγή στατιστικών μετρικών για το εάν η αποτυχία στην αναγνώριση οφείλεται και σε εξωτερικούς παράγοντες, όπως στην ύπαρξη απότομων συναισθηματικών αλλαγών του χρήστη

5 *Βιβλιογραφία*

[1] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, “Deep learning techniques for speech emotion recognition, from databases to models,” *Sensors (Switzerland)*, vol. 21, no. 4, pp. 1–27, 2021, doi: 10.3390/s21041249.

[2] T. He, “EXPLOITING LSTM STRUCTURE IN DEEP NEURAL NETWORKS FOR SPEECH RECOGNITION Key Lab . of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering SpeechLab , Department of Computer Science and Engineering , Shanghai Jiao Tong Unive,” *Icassp 2016*, pp. 5445–5449, 2016.

[3] J. Li, A. Mohamed, G. Zweig, and Y. Gong, “LSTM time and frequency recurrence for automatic speech recognition,” *2015 IEEE Work. Autom. Speech Recognit. Understanding, ASRU 2015 - Proc.*, pp. 187–191, 2016, doi: 10.1109/ASRU.2015.7404793.

[4] W. Yin, K. Kann, M. Yu, and H. Schütze, “Comparative Study of CNN and RNN for Natural Language Processing,” 2017, [Online]. Available: http://arxiv.org/abs/1702.01923.

[5] T. M. Fleming *et al.*, “Serious games and gamification for mental health: Current status and promising directions,” *Front. Psychiatry*, vol. 7, no. JAN, 2017, doi: 10.3389/fpsyt.2016.00215.

[6] P. Sajjadi, A. Ewais, and O. De Troyer, “Individualization in serious games: A systematic review of the literature on the aspects of the players to adapt to,” *Entertain. Comput.*, vol. 41, no. September 2021, p. 100468, 2022, doi: 10.1016/j.entcom.2021.100468.

[7] S. Alves, A. Marques, C. Queirós, and V. Orvalho, “LifeisGame prototype: A serious game about emotions for children with autism spectrum disorders,” *PsychNology J.*, vol. 11, no. 3, pp. 191–211, 2013.

[8] D. N. Y. D. M. P. Prof. Ritu Tiwari Dr. Apoorva Mishra, *Proceedings of International Conference on Computational Intelligence*. 2022.

[9] A. Aksoy, Y. E. Ertürk, S. Erdoğan, E. Eyduran, and M. M. Tariq, “Estimation of honey production in beekeeping enterprises from eastern part of Turkey through some data mining algorithms,” *Pak. J. Zool.*, vol. 50, no. 6, pp. 2199–2207, 2018, doi: 10.17582/journal.pjz/2018.50.6.2199.2207.

[10] F. Sherratt, A. Plummer, and P. Iravani, “Understanding LSTM Network Behaviour of IMU-Based,” *Sensors*, vol. 21, no. 1264, 2021, doi: 10.3390/s21041264.

[11] H. Salman, J. Grover, and T. Shankar, “Hierarchical Reinforcement Learning for Sequencing Behaviors,” vol. 2733, no. March, pp. 2709–2733, 2018, doi: 10.1162/NECO.

[12] F. Saad, H. Mahmud, M. R. Kabir, A. Shaheen, P. Farastu, and K. Hasan, “A Case Study on the Independence of Speech Emotion Recognition in Bangla and English Languages using Language-Independent Prosodic Features.”

[13] S. R. Livingstone and F. A. Russo, *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. 2018.

[14] M. Xu, F. Zhang, and W. Zhang, “Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset,” *IEEE Access*, vol. 9, pp. 74539–74549, 2021, doi: 10.1109/ACCESS.2021.3067460.

[15] M. K. Pichora-Fuller and K. Dupuis, “Toronto emotional speech set (TESS).” Scholars Portal Dataverse, doi: doi:10.5683/SP2/E8H2MF.

[16] Z. K., “University of surrey library,” 2005.

[17] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset,” *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, 2014, doi: 10.1109/TAFFC.2014.2336244.

[18] M. Z. IQBAL, “MFCC and Machine Learning Based Speech Emotion Recognition on,” *Found. Univ. J. Eng. Appl. Sci.*, vol. 1, no. 1, pp. 1–6, 2020.

[19] J. Gudnason and M. Brookes, “Voice source cepstrum coefficients for speaker identification,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. June, pp. 4821–4824, 2008, doi: 10.1109/ICASSP.2008.4518736.

[20] A. Charisma, M. R. Hidayat, and Y. B. Zainal, “Speaker recognition using mel-

frequency cepstrum coefficients and sum square error," in *2017 3rd International Conference on Wireless and Telematics (ICWT)*, 2017, pp. 160–163, doi: 10.1109/ICWT.2017.8284159.

[21] Z. Witold Engel, M. Kłaczyński, and W. Wszołek, "A vibroacoustic model of selected human larynx diseases," *Int. J. Occup. Saf. Ergon.*, vol. 13, no. 4, pp. 367–379, 2007, doi: 10.1080/10803548.2007.11105094.

[22] S. Prasanth, M. Roshni Thanka, E. Bijolin Edwin, and V. Nagaraj, "Speech emotion recognition based on machine learning tactics and algorithms," *Mater. Today Proc.*, no. xxxx, 2021, doi: 10.1016/j.matpr.2020.12.207.

[23] M. C. Bingol and O. Aydogmus, "Implementation of speech recognition for robot control using support vector machine," *Int. Eurasian Conf. Sci. Eng. Technol. (EurasianSciEnTech)*, no. December, pp. 814–819, 2018.

[24] N. Ketkar, "Introduction to Keras," in *Deep Learning with Python: A Hands-on Introduction*, Berkeley, CA: Apress, 2017, pp. 97–111.

[25] D. K. Barupal and O. Fiehn, "Generating the blood exposome database using a comprehensive text mining and database fusion approach," *Environ. Health Perspect.*, vol. 127, no. 9, pp. 2825–2830, 2019, doi: 10.1289/EHP4713.

[26] S. Braun, "LSTM Benchmarks for Deep Learning Frameworks," 2018, [Online]. Available: http://arxiv.org/abs/1806.01818.