

Ατομική Διπλωματική Εργασία

ΣΥΣΤΗΜΑΤΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΝΟΗΜΟΣΥΝΗΣ ΓΙΑ ΔΙΑΓΝΩΣΗ ΣΤΗΝ ΙΑΤΡΙΚΗ :
Η ΠΕΡΙΠΤΩΣΗ ΤΟΥ ΣΥΝΔΡΟΜΟΥ DOWN

ΣΤΑΥΡΟΣ ΚΟΥΛΑΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΑΙΟΣ 2012

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΣΥΣΤΗΜΑΤΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΝΟΗΜΟΣΥΝΗΣ ΓΙΑ ΔΙΑΓΝΩΣΗ ΣΤΗΝ ΙΑΤΡΙΚΗ :
Η ΠΕΡΙΠΤΩΣΗ ΤΟΥ ΣΥΝΔΡΟΜΟΥ DOWN

Σταύρος Κούλας

Επιβλέπων Καθηγητής
Χρίστος Ν. Σχίζας

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου.

Μάιος 2012

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντα καθηγητή κ. Χρίστο Ν. Σχίζα, που με επέλεξε και εμπιστεύτηκε για την εκπόνηση της διπλωματικής αυτής εργασίας. Τον ευχαριστώ ακόμη για την υπομονή αλλά και για την πολύτιμη καθοδήγηση που μου παρείχε.

Επίσης, θα ήθελα να ευχαριστήσω όλους τους συνεργάτες του και ιδιαίτερα τον κ. Κλεάνθη Νεοκλέους, που με τη σειρά του βοήθησε στο μέγιστο με τις ιδέες, την υπομονή αλλά και την καθοδήγησή του και γενικά για την άψογη συνεργασία που είχαμε.

Τέλος, ευχαριστώ όσους με την σειρά τους με στήριξαν και βοήθησαν με το δικό τους τρόπο κατά τη διάρκεια εκπόνησης της διπλωματικής αυτής εργασίας.

Περίληψη

Η εργασία αυτή, έχει σκοπό την χρήση και εφαρμογή των υπολογιστικών μεθόδων που αναπτύχθηκαν, στην προσπάθεια επίλυσης προβλημάτων στην ιατρική και συγκεκριμένα στην πρόβλεψη ιατρικών παθήσεων. Οι μέθοδοι αυτοί, εντάσσονται στο πεδίο της υπολογιστικής νοημοσύνης και συγκεκριμένα στην διπλωματική αυτή εργασία, με νευρωνικά δίκτυα, γενετικούς αλγόριθμους καθώς και με μηχανές υποστήριξης διανυσμάτων.

Η εργασία αυτή ασχολείται με την πρόβλεψη του συνδρόμου Down, γνωστή και ως τρισωμία 21. Ο στόχος είναι να προβλέψουμε έγκαιρα σε όσο το δυνατό μεγαλύτερο βαθμό επιτυχίας κατά πόσο ένα έμβρυο θα αναπτύξει την συγκεκριμένη πάθηση. Η πρόβλεψη αυτή βασίζεται σε πραγματικά δεδομένα που καλύπτουν ένα αριθμό 51208 περιστατικών που είτε ανέπτυξαν είτε όχι την συγκεκριμένη ασθένεια.

Αρχικά, μελετήθηκαν διάφορα άλλα προβλήματα προβλέψεων με νευρωνικά δίκτυα για την ανάπτυξη εμπειρίας αλλά και ανάπτυξη ικανοτήτων ορθής μεθοδολογίας.

Επίσης, δοκιμάστηκαν δύο διαφορετικά λογισμικά εργαλεία, για σκοπούς σύγκρισης και επιλογής του καταλληλότερου.

Τέλος, αναπτύχθηκε ένα υβριδικό μοντέλο γενετικού αλγορίθμου – νευρωνικών δικτύων με στόχο να συνδυάσουμε και να εκμεταλλευτούμε τα θετικά και των δύο αλγορίθμων βελτιώνοντας τα ποσοστά ορθών προβλέψεων σε σύγκριση με άλλες προσεγγίσεις.

Περιεχόμενα

Κεφάλαιο 1 Εισαγωγή - Θεωρητικό Υπόβαθρο

- 1.1 Υπολογιστική Νοημοσύνη
- 1.2 Νευρωνικά Δίκτυα
- 1.3 Μηχανές Υποστήριξης Διανυσμάτων
- 1.4 Γενετικοί Αλγόριθμοι

Κεφάλαιο 2 Σύνδρομο Down

- 2.1 Σύντομη περιγραφή προβλήματος
- 2.2 Ανάλυση δεδομένων
- 2.3 Προηγούμενες προσπάθειες για πρόβλεψη τρισωμιών

Κεφάλαιο 3 Μεθοδολογία

- 3.1 Λογισμικά εργαλεία που χρησιμοποιήθηκαν
 - 3.1.1 Neuroshell2
 - 3.1.2 Matlab
- 3.2 Άλλα προβλήματα προβλέψεων που μελετήθηκαν
- 3.3 Αλγόριθμοι-Τεχνικές που ακολουθήθηκαν
 - 3.3.1 Αλγόριθμος πίσω διάδοσης λάθους
 - 3.3.2 Μηχανές Υποστήριξης Διανυσμάτων
 - 3.3.3 Υβριδικό Μοντέλο Γενετικού Αλγόριθμου – Νευρωνικών Δικτύων

Κεφάλαιο 4 Αποτελέσματα

4.1 Μηχανές Υποστήριξης Διανυσμάτων

4.1.1 Προσπάθεια εξακρίβωσης προβληματικών περιπτώσεων

4.2 Υβριδικό Μοντέλο Γενετικού Αλγόριθμου – Νευρωνικών Δικτύων

Κεφάλαιο 5 Συμπεράσματα – Μελλοντική Εργασία

5.1 Συμπεράσματα διπλωματικής εργασίας

Κεφάλαιο 6 Μελλοντική Εργασία

6.1 Προτάσεις για μελλοντική εργασία

Βιβλιογραφία

Κεφάλαιο 1

Εισαγωγή

1.1 Υπολογιστική Νοημοσύνη	7
1.2 Νευρωνικά Δίκτυα	8
1.3 Μηχανές Υποστήριξης Διανυσμάτων	12
1.4 Γενετικοί Αλγόριθμοι	14

1.1 Υπολογιστική Νοημοσύνη

Ο όρος Υπολογιστική Νοημοσύνη, υποδηλώνει κάθε είδος νοημοσύνης που μπορεί να εκδηλωθεί με υπολογιστικές διαδικασίες.

Ο όρος Τεχνητή Νοημοσύνη αν και χρησιμοποιείται με την ίδια έννοια, είναι πιο γενικός υπό την έννοια ότι υποδηλώνει κάθε είδος νοημοσύνης που δεν είναι βιολογική.

Υπάρχουν δύο προσεγγίσεις Τεχνητής νοημοσύνης. Η ασθενής προσέγγιση, στην οποία πιστεύεται ότι η ανθρώπινη γνώση/νόηση ή ειδικές εκφάνσεις αυτής μπορεί να προσομοιαστούν σε υπολογιστικά συστήματα, και η ισχυρή προσέγγιση, στην οποία

πιστεύεται ότι η ανθρώπινη γνώση/νόηση μπορεί να εκδηλωθεί αυτόματα σε υπολογιστικά συστήματα.

Στην υπολογιστική νοημοσύνη, η βασική ιδέα-προσέγγιση είναι αντί να αναπαριστούμε την γνώση σε ένα υψηλό επίπεδο, αντικαθιστούμε τα μοντέλα αυτά με αλγόριθμους που στηρίζονται στην αριθμητική επεξεργασία της πληροφορίας (δηλαδή υπολογιστικά). Έτσι, αντί να έχουμε ένα αναλυτικό μαθηματικό μοντέλο για την εφαρμογή, κατασκευάζουμε έναν αλγόριθμο ο οποίος αναλύει και μαθαίνει από δεδομένα που ήδη υπάρχουν και περιγράφουν την εφαρμογή [1].

Μερικές μέθοδοι μελέτης της υπολογιστικής νοημοσύνης είναι :

- Τεχνητά νευρωνικά δίκτυα
- Γενετικοί Αλγόριθμοι
- Ασαφής Λογική
- Συστήματα Τεχνητής Ζωής

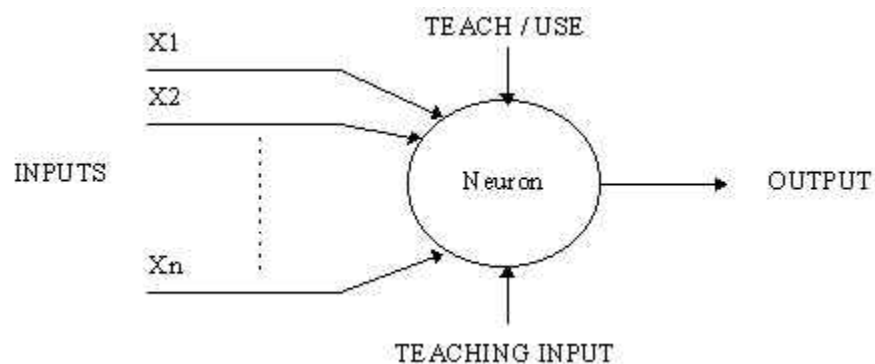
1.2 Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα είναι ένα είδος τεχνητής νοημοσύνης, εμπνευσμένο τον τρόπο με τον οποίο δουλεύει ο ανθρώπινος εγκέφαλος. Αποτελείται από ένα μεγάλο αριθμό διασυνδεδεμένων στοιχείων -τους νευρώνες- οι οποίοι συνεργάζονται για την επίλυση συγκεκριμένων προβλημάτων. Όπως και με τον ανθρώπινο εγκέφαλο, τα τεχνητά νευρωνικά δίκτυα μαθαίνουν από παραδείγματα. Αυτό επιτυγχάνεται όπως και στον ανθρώπινο εγκέφαλο με προσαρμογές των συναπτικών συνδέσεων που υπάρχουν μεταξύ των νευρώνων. Κάθε δίκτυο, μπορεί να τροποποιηθεί αναλόγως του προβλήματος. Τέτοια παραδείγματα είναι προβλήματα ταξινόμησης όπως και αυτό του συνδρόμου down.

Ο τρόπος με τον οποίο τα νευρωνικά δίκτυα μαθαίνουν και είναι ικανά να διαχωρίσουν δεδομένα είναι ο εξής :

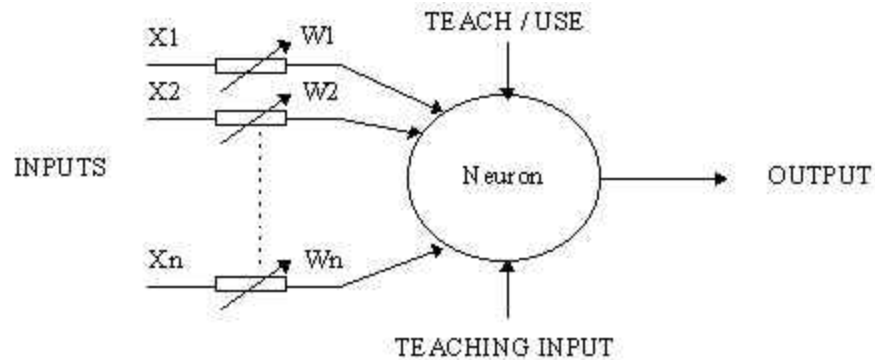
Χρησιμοποιούμε δεδομένα (εισόδους), για τα οποία γνωρίζουμε την έξοδο. Για παράδειγμα για την περίπτωση του συνδρόμου down, ιατρικά δεδομένα και κατά πόσο στην κάθε περίπτωση το βρέφος (διότι είναι σημαντικό να προγνωστεί έγκαιρα), στο μέλλον ανέπτυξε την συγκεκριμένη νόσο.

Έτσι, το δίκτυο με την παρουσίαση μιας νέας περίπτωσης, δηλαδή νέων ιατρικών δεδομένων, μπορεί να βρει τα κοινά χαρακτηριστικά και να γενικεύσει δίνοντας μας έγκυρο αποτέλεσμα σε πολύ μικρό χρονικό διάστημα και έτσι ουσιαστικά το δίκτυο αποκτά εμπειρία.



Σχήμα 1.2.1 Ένας απλός νευρώνας [6]

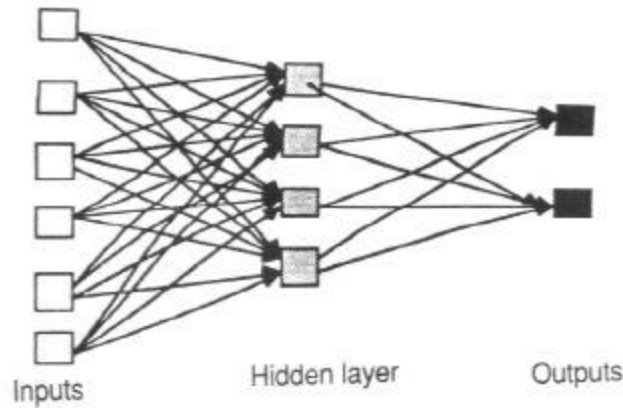
Το πρώτο τεχνητό νευρωνικό δίκτυο (Σχήμα 1.2.2) προτάθηκε από τους Warren McCulloch και Walter Pitts το 1943 [2]. Με την πάροδο του χρόνου πολλοί επιστήμονες ανέπτυξαν το μοντέλο MCP το οποίο πήρε το όνομα του από τους δύο αυτούς επιστήμονες (McCulloch Pitts Neurons), και το επόμενο μεγάλο βήμα εισηχθη από τον Frank Rosenblatt το 1958 [4]. Πρότεινε ένα νέο μοντέλο -το perceptron- που ουσιαστικά είναι ένας νευρώνας MCP, όπου οι είσοδοι πρώτα περνούν από κάποιους προεπεξεργαστές. Αυτό επέτρεπε λύση σε προβλήματα όπου η είσοδος θα είχε ως έξοδο 2 πιθανά αποτελέσματα [5].



Σχήμα 1.2.2 Ένας MCP νευρώνας [6]

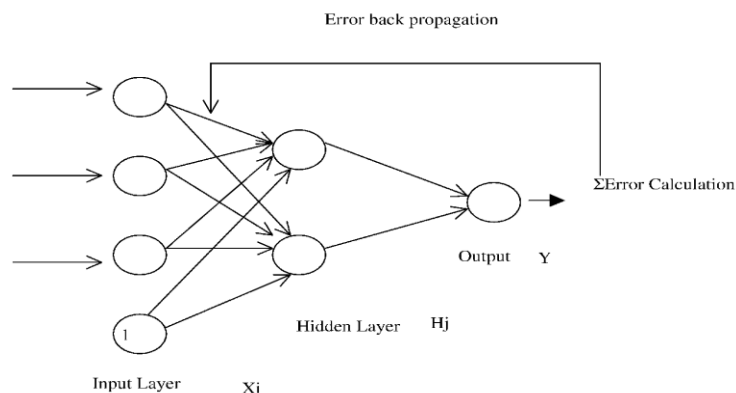
Η διαφορά με το προηγούμενο παράδειγμα απλού νευρώνα, είναι ότι σε αυτό το μοντέλο οι εισόδους είναι συνδεδεμένοι με βάρη ($W_1, W_2, \dots, W_n$), που έχει αποτέλεσμα ότι για κάθε είσοδο το αποτέλεσμα εξαρτάται και από την τιμή του βάρους, το οποίο είναι ουσιαστικά ένας αριθμός που πολλαπλασιάζεται με την τιμή εισόδου. Όταν το άθροισμα όλων των εισόδων (πολλαπλασιασμένο με το βάρος τους) ξεπεράσει μια συγκεκριμένη τιμή -κατώφλι-, τότε ο νευρώνας λέμε ότι πυροβολεί. Οπότε μπορεί να γίνει διαχωρισμός των εξόδων μας σε δύο περιπτώσεις, τις περιπτώσεις που ο νευρώνας πυροβολεί ή όχι. Η δυνατότητα προσαρμοσης των βαρών και της τιμής κατωφλίου, καθιστά τα τεχνητά νευρωνικά δίκτυα ευέλικτα και κατάλληλα για πολλά διαφορετικά προβλήματα.

Υπάρχουν επίσης, τέτοια δίκτυα τα οποία χρησιμοποιούν περισσότερα επίπεδα από αυτά της εισόδου και εξόδου, τα κρυφά επίπεδα, τα οποία μπορεί να είναι είτε ένα είτε δύο και παίρνουν σαν είσοδο τις τιμές των εισόδων αφού πολλαπλασιαστούν με τα βάρη. Ο κάθε κόμβος σε αυτό το επίπεδο έχει επίσης το δικό του βάρος που με τη σειρά του πολλαπλασιάζεται και αθροίζεται με όλα του κρυφού επιπέδου και έπειτα περνά στο επίπεδο εξόδου (Σχήμα 1.2.3).



Σχήμα 1.2.3 Παράδειγμα απλού δικτύου με κρυφό επίπεδο [6]

Ένα πιο εξελιγμένο μοντέλο, το οποίο χρησιμοποιήθηκε στα πλαίσια της διπλωματικής αυτής εργασίας, είναι αυτό του Αλγόριθμου πίσω διάδοσης λάθους (Back Propagation). Το μοντέλο αυτό, προσαρμόζει τα βάρη της κάθε εισόδου με τρόπο ώστε το σφάλμα μεταξύ της επιθυμητής εξόδου (ορθή απάντηση) και της πραγματικής εξόδου (αυτής που υπολογίζεται από το δίκτυο) να είναι όσο το δυνατό πιο μικρό. Αυτό επιτυγχάνεται ουσιαστικά στέλνοντας προς τα πίσω (στο επίπεδο εισόδου) την τιμή που υπολογίστηκε και με βάση την επιθυμητή τιμή προσαρμόζονται τα βάρη, μια διαδικασία η οποία γίνεται συνεχώς στο δίκτυο.



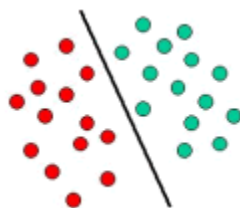
Σχήμα 1.2.3 Παράδειγμα αλλαγής βαρών με Αλγόριθμο πίσω διάδοσης λάθους [6]

Μερικοί παράμετροι των νευρωνικών δικτύων που μπορούν να τροποποιηθούν, είναι ο ρυθμός μάθησης, δηλαδή το πόσο γρήγορα μαθαίνει το δίκτυο, ορμή για να ξεφύγει ο αλγόριθμος από τοπικά ελάχιστα και να αναζητήσει λύση και σε άλλα επίπεδα, συναρτήσεις ενεργοποίησης και κατωφλίου, αριθμός νευρώνων στα κρυφά επίπεδα και πολλές άλλες προσαρμογές που κάνουν τα νευρωνικά δίκτυα ευέλικτα και κατάλληλα για πολλά διαφορετικά προβλήματα.

1.3 Μηχανές Υποστήριξης Διανυσμάτων

Ο αλγόριθμος αυτός αρχικά εφευρέθηκε από τον Vladimir N. Vapnik και το πρότυπο το οποίο ισχύει σήμερα είχε προταθεί από τον ίδιο Vladimir N. Vapnik και την Corinna Cortes το 1995 [26].

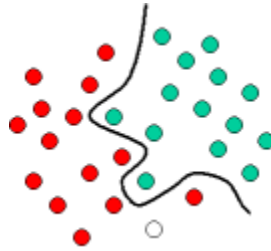
Είναι βασισμένα στην έννοια των decision planes, που καθορίζουν κάποια όρια αποφάσεων. Ένα decision plane, ορίζεται αυτό που διαχωρίζει ένα σύνολο αντικειμένων τα οποία έχουν διαφορετικά χαρακτηριστικά – κλάσεις, όπως για παράδειγμα στην παρακάτω εικόνα όπου τα αντικείμενα χωρίζονται στις κατηγορίες πράσινο – κόκκινο και κάθε νέο αντικείμενο, ανάλογα στην κλάση που υπάγεται (πράσινο - κόκκινο) πρέπει να πέσει στα ανάλογο κομμάτι.



Εικόνα 1.3.2 Linear Separated Problems [3]

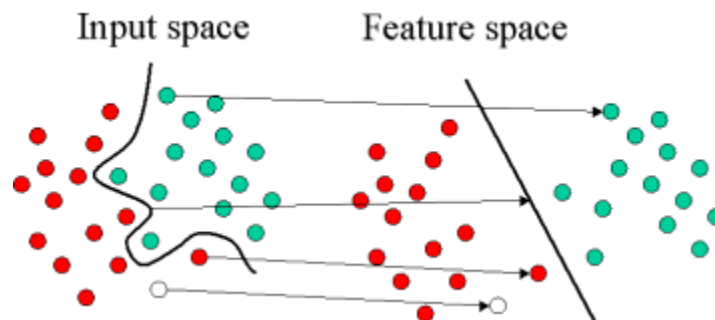
Το πιο πάνω παράδειγμα, αναφέρεται σε προβλήματα γραμμικά επιλύσιμα, που μπορούν δηλαδή να διαχωριστούν με μια ευθεία. Τα περισσότερα προβλήματα όμως δεν είναι τόσο απλά (βλ. εικόνα 1.3.2) και συχνά χρειάζονται αρκετά περίπλοκες δομές προκειμένου να γίνει ο βέλτιστος διαχωρισμός, έτσι ώστε βάση των περιπτώσεων στις οποίες εκπαιδεύσαμε τον αλγόριθμο, να διαχωρίσει ορθά τις καινούργιες-άγνωστες

περιπτώσεις, και εδώ έρχεται το μεγαλύτερο πλεονέκτημα των μηχανών υποστήριξης διανυσμάτων, ότι είναι ιδιαίτερα αποδοτικές για τέτοιας φύσης προβλήματα.



Εικόνα 1.3.2 Non – Linear Separated Problems [3]

Στην ακόλουθη εικόνα (Εικόνα 1.3.3), φαίνεται η βασική ιδέα των Μηχανών Υποστήριξης Διανυσμάτων. Τα αρχικά διαχωρισμένα αντικείμενα, αναδιατάσσονται με μια σειρά από μαθηματικές πράξεις γνωστές ως kernels ($K(x,y)$). Η διαδικασία αυτή, είναι γνωστή και ως χαρτογράφηση (mapping) και παρατηρούμε ότι μετά την αναδιάταξη (στα δεξιά της εικόνας), τα αντικείμενα είναι γραμμικά διαχωρίσιμα και επομένως αντί για διαχωρισμό των αντικειμένων με καμπύλη (στα αριστερά της εικόνας), απλά απομένει να βρούμε την βέλτιστη γραμμή που μπορεί να διαχωρίσει τα αντικείμενα μεταξύ τους.



Εικόνα 1.3.3 Re-arrangement of data space [3]

Ο αλγόριθμος αυτός, επιδέχεται κάποιων παραμέτρων, μεταξύ των οποίων και ένα error-penalty, το οποίο καθορίζεται από τον προγραμματιστή και έχει ως στόχο να αγνοήσει κάποια αντικείμενα τα οποία θα οδηγούσαν στην δραματική αλλαγή της

γραμμής διαχωρισμού έτσι ώστε να διαχωρίζει ακριβώς όλα ή τα περισσότερα αντικείμενα, κάτι το οποίο θα έχει ως αποτέλεσμα την υπερμάθηση, δηλαδή δίνει βαρύτητα στο να μαθαίνει τις λεπτομέρειες των προτύπων πάνω στα δεδομένα που εκπαιδεύεται, και έτσι χάνει την ικανότητα να γενικεύει και ουσιαστικά να αδυνατεί να κατατάξει ένα καινούργιο-άγνωστο αντικείμενο. Επομένως, η επιλογή του ορθού error, είναι και η πρόκληση του αλγορίθμου αυτού, κάτι που μπορεί να προσδιοριστεί μέσα από πειραματική μελέτη των αποτελεσμάτων.

1.3 Γενετικοί Αλγόριθμοι

Οι Γενετικοί Αλγόριθμοι ανήκουν στην κατηγορία των εξελικτικών αλγόριθμων, οι οποίοι βασίζονται στις αρχές της βιολογικής εξέλιξης.

Υλοποιήθηκαν από τον John Henry Holland το 1975 σε bit-strings [28], όμως σήμερα υπάρχουν εφαρμογές σε γράφους, διανύσματα πραγματικών αριθμών και διατεταγμένες λίστες.

Οι Γενετικοί Αλγόριθμοι χαρακτηρίζονται από την απλότητα και την κομψότητά τους ως γεροί αλγόριθμοι αναζήτησης, καθώς επίσης και από τη δύναμή τους να ανακαλύπτουν γρήγορα τις καλές λύσεις δύσκολων, μεγάλης διάστασης προβλημάτων.

Γενικά, βασίζονται στη μέθοδο του τυχαίου ψαξίματος με τη μόνη διαφορά ότι το χάσιμο χρόνου μέχρι να βρεθεί η λύση μπορεί να αποφευχθεί χρησιμοποιώντας μεθόδους βασισμένες στην βιολογία και είναι ιδιαίτερα χρήσιμοι και αποδοτικοί όταν:

- Το διάστημα αναζήτησης είναι μεγάλο, σύνθετο ή ανεπαρκώς κατανοητό.
- Η γνώση περιοχών είναι λιγοστή ή η ειδική γνώση είναι δύσκολο να κωδικοποιηθεί ώστε, να περιοριστεί στο διάστημα αναζήτησης.
- Καμία μαθηματική ανάλυση δεν είναι διαθέσιμη.
- Οι παραδοσιακές μέθοδοι αναζήτησης αποτυγχάνουν.

Τα βήματα εκτέλεσης ενός γενετικού αλγόριθμου εφαρμόζονται πάνω σε ένα πληθυσμό πιθανών λύσεων και εξελίσσουν τον πληθυσμό εφαρμόζοντας σε αυτόν διάφορες γενετικές διαδικασίες, όπως διαδικασίες επιλογής, αναπαραγωγής, μετάλλαξης. Ακολουθώντας δημιουργούν νέο πληθυσμό –με τους καλύτερους “απογόνους”- που αντικαθιστά τον προηγούμενο.

Συγκρινόμενοι με άλλους παραδοσιακούς αλγόριθμους βελτιστοποίησης, οι γενετικοί αλγόριθμοι δουλεύουν με κωδικοποιημένες παραμέτρους, δεν ψάχνουν μόνο σε ένα σημείο αλλά από ένα πλήθος σημείων, χρησιμοποιούν πληροφορίες που οι ίδιοι παράγουν αντί να είναι εξωτερικά παραγόμενες ή βοηθητικές γνώσεις και χρησιμοποιούν όχι ντετερμινιστικούς αλλά πιθανοθεωρητικούς κανόνες.

Μερικά από τα πλεονεκτήματα τους είναι ότι μπορούν να συμμετέχουν σε υβριδικές μορφές με άλλες μεθόδους, έχουν από τη φύση τους το στοιχείο του παραλληλισμού, είναι εύκολα επεκτάσιμοι και εξελίξιμοι και μπορούν να επιλύουν δύσκολα προβλήματα γρήγορα και αξιόπιστα [1].

Ένας απλός Γενετικός Αλγόριθμος αποτελείται από τα ακόλουθα βήματα :

1. Κωδικοποίηση (Coding)
2. Αρχικοποίηση (Initialization)
3. Αποκωδικοποίηση (Decoding)
4. Υπολογισμός ικανότητας ή αξιολόγηση (Fitness calculation ή evaluation)
5. Επιλογή (Selection)
6. Αναπαραγωγή (Reproduction)
7. Διασταύρωση (Crossover ή mating)
8. Μετάλλαξη (Mutation)
9. Επανάληψη από το βήμα (2) μέχρι να ικανοποιηθεί το κριτήριο τερματισμού του Γ.Α.

Αναλυτικότερα για το κάθε βήμα :

- Βήμα 1: Γίνεται μια κωδικοποίηση του συνόλου τιμών που μπορούν να λάβουν οι μεταβλητές που θα χρησιμοποιηθούν. Επίσης καθορίζεται το μέγεθος του πληθυσμού, δηλαδή το πλήθος των διαφορετικών χρωμοσωμάτων πάνω στα οποία θα εκτελείται ο αλγόριθμος.
- Βήμα 2 : Οι τιμές των μεταβλητών για κάθε χρωμόσωμα αρχικοποιούνται με τυχαίες τιμές στο εύρος που καθορίζεται στο Βήμα 1.
- Βήμα 3 : Γίνεται αποκωδικοποίηση των τιμών των παραμέτρων από δυαδική σε δεκαδική μορφή, για να μπορούν να τύχουν περαιτέρω επεξεργασίας.
- Βήμα 4 : Αποτίμηση της ικανότητας κάθε χρωμοσώματος βάση της συνάρτησης αξιολόγησης.
- Βήμα 5 : Επιλογή χρωμοσωμάτων με την καλύτερη αξιολόγηση που λειτουργούν ως γονείς για το επόμενο βήμα.
- Βήμα 6 & 7 Σύμφωνα με προκαθορισμένη πιθανότητα διασταύρωσης, σχηματίζονται απόγονοι μετά από την διασταύρωση των χρωμοσωμάτων των γονέων (ζευγάρι). Αν δεν εφαρμοστεί η διασταύρωση οι απόγονοι είναι μια πιστή αντιγραφή των γονέων τους.
- Βήμα 8 : Στο βήμα αυτό γίνεται τυχαία αλλαγή σε κάποια θέση του χρωμοσώματος.

Κεφάλαιο 2

Σύνδρομο Down

2.1 Σύντομη Περιγραφή προβλήματος	17
2.2 Ανάλυση Δεδομένων	19
2.3 Προηγούμενες προσπάθειες για πρόβλεψη τρισωμιών	21

2.1 Σύντομη Περιγραφή Προβλήματος

Το σύνδρομο Down είναι από τις πιο συχνές τρισωμίες που παρουσιάζονται και περίπου μία στις 700 γεννήσεις παιδιών επηρεάζεται από αυτό το σύνδρομο.

Αρχικά αναγνωρίστηκε το 1866 από το γιατρό Langdon Down [30], ο οποίος πρόσεξε ότι μια μικρή ομάδα ατόμων, άσχετων μεταξύ τους, που βρίσκονταν σε ιδρύματα είχαν παρόμοια σωματικά χαρακτηριστικά. Το 1958, ανακαλύφθηκε η αιτιολογία των σωματικών χαρακτηριστικών και των μαθησιακών δυσκολιών από το γιατρό-νηπιαγωγό Jerome Lejeune στο Παρίσι [31]. Η αιτιολογία ήταν η παρουσία ενός παραπανίσιου χρωμοσώματος σε κάθε κύτταρο του σώματος.

Ένας φυσιολογικός άνθρωπος γεννιέται με 46 χρωμοσώματα, τα οποία κληρονομεί σε ίσο αριθμό (από 23) από τον κάθε γονέα. Τα χρωμοσώματα αυτά, καθορίζουν ανθρώπινα χαρακτηριστικά όπως για παράδειγμα το χρωμοσώματα 23, που καθορίζει το φύλο. Για το συγκεκριμένο χρωμόσωμα, η μητέρα δίνει πάντα X και ανάλογα με το

χρωμόσωμα που θα κληρονομηθεί από τον πατέρα θα προκύψει το φύλο του παιδιού. Για παράδειγμα αν ο πατέρας δώσει X θα έχουμε XX στο χρωμόσωμα 23 και άρα κορίτσι, ενώ στην αντίθετη περίπτωση που ο πατέρας δώσει Y θα έχουμε XY και άρα αγόρι.

Σε περίπτωση που το έμβρυο αντί για 46 χρωμοσώματα, έχει 47, τότε μιλάμε για γενετική διαταραχή – τρισωμία. Στην περίπτωση συνδρόμου Down - τρισωμία 21, το χρωμόσωμα που παρουσιάζεται 3 φορές, είναι το χρωμόσωμα 21.

Υπάρχουν πολλές τέτοιες γενετικές διαταραχές – τρισωμίες όπως για παράδειγμα η τρισωμία 13 (Patau Syndrome) και η τρισωμία 18 (Edward' s Syndrome).

Μερικά χαρακτηριστικά παιδιού με σύνδρομο Down [24], είναι :

- Παρορμητική συμπεριφορά
- Κακή Κρίση
- Διαλείπουσα Προσοχή
- Αργή Μάθηση
- Αγανάκτηση και θυμός λόγω επίγνωσης της κατάστασης – ορίων τους
- Καρδιακά προβλήματα (π.χ. έλλειμμα μεσοκοιλιακού διαφράγματος)
- Επηηρεασμός όρασης (π.χ. καταρράκτης)
- Ακουστικά προβλήματα
- Υποθυρεοειδισμός

Στόχος είναι να προβλέψουμε την πιθανότητα εμφάνισης του συνδρόμου το συντομότερο δυνατό, έτσι ώστε να ενημερωθούν οι γονείς και να διαχειριστούν όπως οι ίδιοι επιθυμούν την εγκυμοσύνη.

Μέχρι σήμερα, οι τρόποι διάγνωσης γίνονταν με βάση την εμφάνιση του βρέφους, καθώς υπάρχουν κάποια σωματικά χαρακτηριστικά που διακρίνουν ένα έμβρυο με σύνδρομο Down, όπως για παράδειγμα μικρό μέγεθος χείλους και αυτιών, περίσσεια του δέρματος στον αυχένα του λαιμού, πλατιά χέρια με μικρά δάκτυλα. Επίσης μέσω

διάφορων εξετάσεων όπως για παράδειγμα ακτινογραφίες του στήθους και της γαστρεντερικής οδού και υπερηχοκαρδιογράφημα για ανωμαλίες της καρδιάς.

Στόχος της διπλωματικής αυτής εργασίας, είναι να βοηθήσει στην πρόβλεψη τέτοιων παθήσεων, στην συγκεκριμένη περίπτωση χρησιμοποιώντας πληθώρα χαρακτηριστικών που έχουν αποδειχθεί να παίζουν σημαντικό ρόλο στην εμφάνιση ή όχι του συνδρόμου down όπως για παράδειγμα :

- Ηλικία μητέρας
- Κληρονομικότητα
- Μέγεθος ρινικού οστού
- Tricuspid flow
- Ductus venosus flow

2.2 Ανάλυση Δεδομένων

Τα δεδομένα στα οποία βασίστηκε η διπλωματική αυτή εργασία, είναι μια μεγάλη συλλογή δεδομένων από το Fetal Medicine Foundation στο Λονδίνο, τα οποία αφορούν διάφορες μετρήσεις από περιπτώσεις εγκυμονών γυναικών, οι οποίες καθορίζουν σε μεγάλο βαθμό αν τα βρέφη είτε θα επιδείξουν είτε όχι το σύνδρομο Down. Συγκεκριμένα, τα δεδομένα αποτελούνται από 51208 περιπτώσεις από τις οποίες οι 691 αφορούν περίπτωση εμφάνισης συνδρόμου down.

Έτσι, μπόρεσα να αναπτύξω και να εκπαιδεύσω διάφορα συστήματα και να ελέγξω την ορθότητά τους.

Οι μετρήσεις – είσοδοι στις οποίες στηρίχθηκε η διπλωματική αυτή εργασία είναι οι παρακάτω :

- **Mothers Age**

Στην περίπτωση αυτή έχουμε την ηλικία της μητέρας η οποία παίζει καθοριστικό ρόλο, και συγκεκριμένα όσο μεγαλύτερη είναι, τόσο μεγαλύτερες είναι οι πιθανότητες το βρέφος να αναπτύξει το σύνδρομο Down, ειδικά μετά την ηλικία των 30.

- **Previous T21, T18, T13**

Κατά πόσο η γυναίκα είχε στο παρελθόν εγκυμοσύνες όπου το παιδί ανέπτυξε γενετική ανωμαλία. Συγκεκριμένα 3 διαφορετικές στήλες δεδομένων για την κάθε τρισωμία.

- **Crown Rump Length (CRL)**

Μέτρηση που αφορά το μήκος του κεφαλουραίου του βρέφους.

- **Pregnancy Associated Plasma Protein A (PAPP-A MoM)**

Ποσότητα πρωτεΐνης Α, της οποίας μικρές ποσότητες κρίνονται ως προδιάθεση εμφάνισης γενετικών ανωμαλιών [7].

- **Human Chorionic Gonadotropin (B-Hcg MoM)**

Ποσότητα χοριακής γοναδοτροπίνης. Αφορά ορμόνη η οποία παράγεται κατά την διάρκεια της εγκυμοσύνης.

- **Nasal Bone**

Κατά πόσο έχει σχηματιστεί ορθά το ρινικό οστό σε κάποιο συγκεκριμένο χρονικό διάστημα (11-13 εβδομάδες + 6 μέρες) [8].

- **Tricuspid flow**

Εκτίμηση της τριγλώχινος ροής [8].

- **Ductus venosus flow**

Εκτίμηση ροής Φλεβώδη πόρου [8].

- **Differential Delta by Nuchal Translucency (Delta NT)**

Μετρήσεις απόκλισης από υπερηχογραφήματα [9].

2.3 Προηγούμενες Προσπάθειες για Πρόβλεψη Τρισωμιών

Η πρόκληση της διπλωματικής αυτής εργασίας ήταν όχι μόνο να πετύχουμε όσο το δυνατό μεγαλύτερο ποσοστό επιτυχίας αλλά και να δοκιμαστούν κάποιοι νέοι μέθοδοι-προσεγγίσεις, πάντα στα πλαίσια της υπολογιστικής νοημοσύνης. Αποτελέσματα από προηγούμενες προσπάθειες πρόβλεψης για το σύνδρομο Down φαίνονται πιο κάτω.

Προσπάθεια 1 [11] :

Total number of cases	34527
Normal cases	34016
T21 cases	511
Training set cases	26555
Training set, normal cases	26147
Training set, T21 cases	408
Verification set cases	7941
Verification set, normal cases	7850
Verification set, T21 cases	91

Αποτελέσματα :

SET	SIZE	Προσπάθεια 1	Προσπάθεια 2	Προσπάθεια 3
Training set correct cases	34527	57%	71%	75.5 %
Verification set correct cases	92%	54%	59%	72.5%

Επίσης, σε μια άλλη προσπάθεια πρόβλεψης χρησιμοποιώντας νευρωνικά δίκτυα με επίσης διαφορετικό αρχείο εισόδου, προέκυψαν τα πιο κάτω αποτελέσματα [29].

Total number of cases	23513
Normal cases	23296
T21 cases	217
Training set cases	23311
Training set, normal cases	23134
Training set, T21 cases	177
Test set cases	109
Test set, normal cases	84
Test set, T21 cases	25
Verification set cases	93
Verification set, normal cases	78
Verification set, T21 cases	15

Αποτελέσματα :

SET	SIZE	% of total cases
Training set correct cases	23307	99.9
Training set T21 correct cases	175	98.9
Test set correct cases	109	100.0
Training set T21 correct cases	25	100.0
Verification set correct cases	92	98.9
Verification set T21 correct cases	14	93.3

Ωστόσο, και στις δύο περιπτώσεις, χρησιμοποιήθηκαν λιγότερα δεδομένα και συγκεκριμένα μικρότερο δείγμα περιπτώσεων επίσης από το Fetal Medicine Foundation.

Κεφάλαιο 3

Μεθοδολογία

3.1 Λογισμικά εργαλεία που χρησιμοποιήθηκαν	23
3.1.1 Neuroshell2	25
3.1.2 Matlab	27
3.2 Άλλα προβλήματα προβλέψεων που μελετήθηκαν	27
3.3 Αλγόριθμοι-Τεχνικές που ακολουθήθηκαν	32
3.3.1 Αλγόριθμος πίσω διάδοσης λάθους	32
3.3.2 Μηχανές Υποστήριξης Διανυσμάτων	35
3.3.3 Υβριδικό Μοντέλο Γενετικού Αλγόριθμου – Νευρωνικών Δικτύων	36

3.1 Λογισμικά εργαλεία που χρησιμοποιήθηκαν

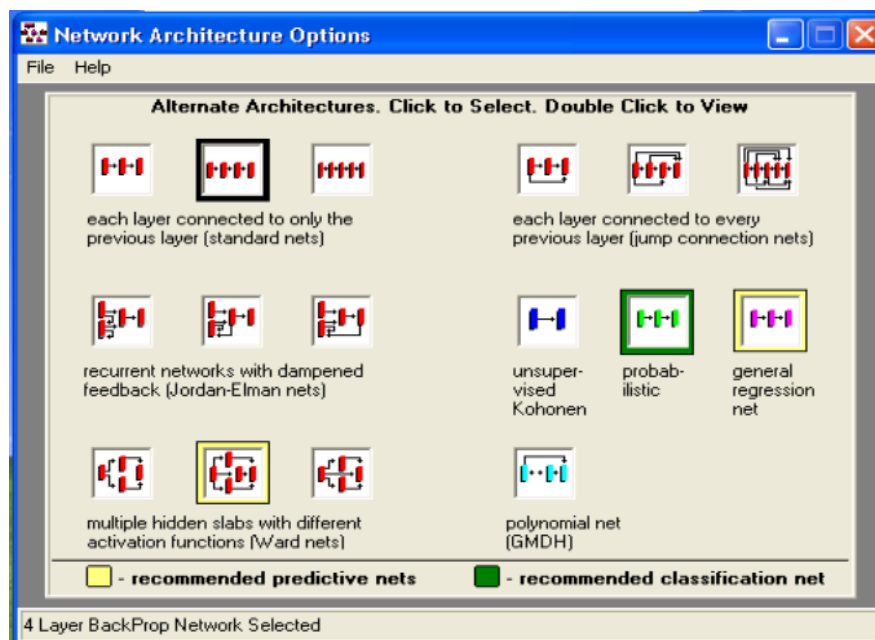
Στα πλαίσια της διπλωματικής αυτής εργασίας, έχω χρησιμοποιήσει το λογισμικό εργαλείο Neuroshell2, το οποίο έχει υλοποιημένους προγραμματιστικά αλγόριθμους νευρωνικών δικτύων.

Επίσης, χρησιμοποίησα την γλώσσα προγραμματισμού Matlab, στην οποία έχω υλοποιήσει αλγόριθμο πίσω διάδοσης λάθους σε μια προσπάθεια σύγκρισης των δύο και επιλογής του καταλληλότερου για το πρόβλημα πρόβλεψης της τρισωμίας 21 – συνδρόμου Down.

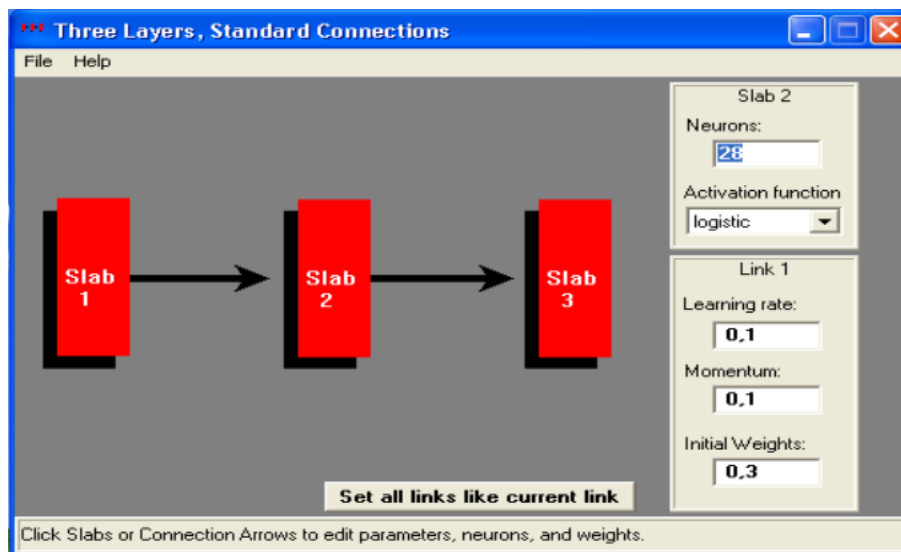
Τέλος, έχω υλοποιήσει ένα υβριδικό μοντέλο γενετικού αλγόριθμου σε συνδυασμό με νευρωνικά δίκτυα στην γλώσσα προγραμματισμού Matlab, το οποίο θα αναφερθεί αναλυτικά στην συνέχεια.

3.1.1 Neuroshell2

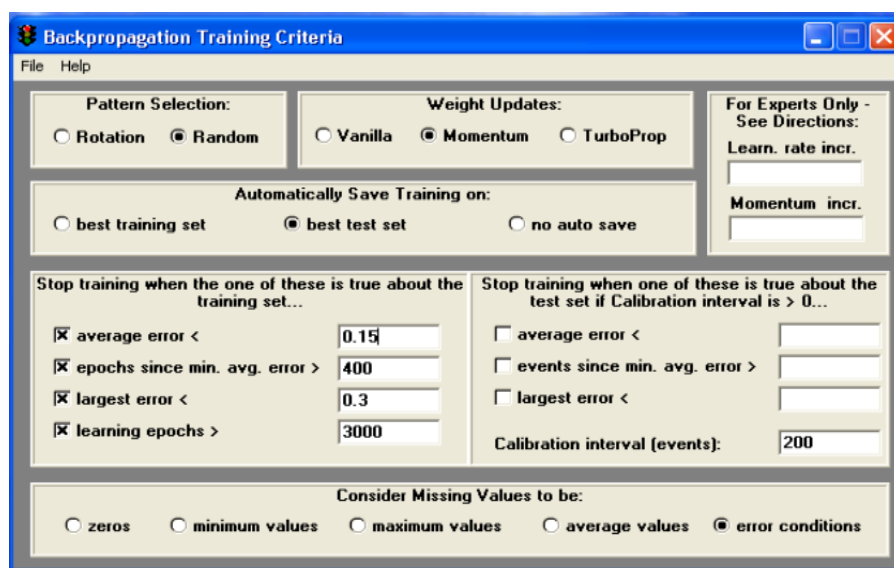
Το λογισμικό αυτό, για το οποίο μου δόθηκε άδεια χρήσης από το Πανεπιστήμιο Κύπρου, όπως έχω προαναφέρει προσφέρει δυνατότητα επιλογής δικτύου, υποστηρίζοντας αρκετές διαφορετικές αρχιτεκτονικές, καθώς και την δυνατότητα διαμόρφωσης των παραμέτρων του δικτύου, όπως για παράδειγμα ρυθμού μάθησης, εποχών επανάληψης, ορμή, συναρτήσεις ενεργοποίησης αλλά και πολλές άλλες προσαρμογές όπως φαίνονται στα πιο κάτω σχήματα [10].



Σχήμα 3.1.1.1 Επιλογή αρχιτεκτονικής



Σχήμα 3.1.1.2 Καθορισμός Παραμέτρων Αρχιτεκτονικής



Σχήμα 3.1.1.3 Περισσότερες επιλογές παραμετροποίησης

Αρκεί να φορτωθεί στο λογισμικό αρχείο τύπου .csv με τα δεδομένα εισόδου και ακολούθως δίνονται στον χρήστη επιλογές ως προς τα δεδομένα εισόδου, για παράδειγμα να αγνοήσει μερικές στήλες, υπολογισμός απόκλισης και πολλές άλλες επιλογές αρκετά εύκολα και γρήγορα.

3.1.2 Matlab

Πέρα από τις πολλές δυνατότητες του λογισμικού Neuroshell2, επέλεξα να υλοποιήσω υπολογιστικά συστήματα και να επικεντρωθώ περισσότερο στην Matlab, καθώς είχαν γίνει προσπάθειες στο παρελθόν για το συγκεκριμένο πρόβλημα με το λογισμικό αυτό [11], και επίσης στην Matlab είχα την δυνατότητα να αναπτύξω τον δικό μου κώδικα και να μπορώ να τροποποιήσω με αυτό τον τρόπο όλες ανεξαρτήτως τις παραμέτρους.

Συγκεκριμένα, χρησιμοποίησα νευρωνικό δίκτυο με Αλγόριθμο πίσω διάδοσης λάθους ενός επιπέδου. Χρησιμοποίησα 100 επαναλήψεις – εποχές με συνάρτηση μάθησης για το νευρωνικό δίκτυο την συνάρτηση `traingdm()` της Matlab η οποία ανανεώνει τις τιμές των βαρών και του bias σύμφωνα με την μεθοδο κατάβασης κλίσης με ορμή. Σαν κανόνα αλλαγής των βαρών και του bias χρησιμοποίησα την συνάρτηση `'adaptwb'` η οποία πάντα ακολουθείται από την `'learnp'` (κανόνας αλλαγής Perceptron). Για υπολογισμό της απόδοσης του δικτύου, χρησιμοποίησα την `'mse'` (mean squared error) η οποία δεδομένων των πραγματικών αποτελεσμάτων, του επιθυμητού αποτελέσματος, καθώς και το σφάλμα στις τιμές των βαρών χρησιμοποιεί για υπολογισμό τον μέσο όρων των τετραγώνων των τιμών των σφαλμάτων [21].

3.2 Άλλα προβλήματα προβλέψεων που μελετήθηκαν

Στην προσπάθεια να κατανοήσω καλύτερα τα νευρωνικά δίκτυα και πως μπορώ για ένα πρόβλημα να βρω την βέλτιστη λύση, όσο αφορά το ποσοστό επιτυχίας πρόβλεψης, πέρα από την θεωρία άρχισα να κατασκευάζω δίκτυα, χρησιμοποιώντας διάφορα γνωστά παραδείγματα κατηγοριοποίησης [12], προσπαθώντας να βρω μια λύση η οποία να είναι η καλύτερη ή ισάξια με τις μέχρι τώρα λύσεις οι οποίες έχουν προταθεί.

Ένα από τα προβλήματα αυτά, είναι το πρόβλημα πρόβλεψης καρκίνου του μαστού.

Χρησιμοποίησα ένα γνωστό σύνολο δεδομένων, το οποίο περιέχει μετρήσεις από περιστατικά στα οποία το βρέφος ανέπτυξε ή όχι καρκίνο του μαστού. Συνολικά, το αρχείο αποτελείται από παραδείγματα 699 περιστατικών για τα οποία υπάρχουν

μετρήσεις που αφορούν τα ακόλουθα σωματικά χαρακτηριστικά και το εύρος τιμών τους.

- Clump Thickness 1 - 10
- Uniformity of Cell Size 1 - 10
- Uniformity of Cell Shape 1 - 10
- Marginal Adhesion 1 - 10
- Single Epithelial Cell Size 1 - 10
- Bare Nuclei 1 - 10
- Bland Chromatin 1 - 10
- Normal Nucleoli 1 - 10
- Mitoses 1 - 10

Μετά από πολλές προσπάθειες διάφορων αρχιτεκτονικών και παραμέτρων, κατέληξα στα ακόλουθα-καλύτερα αποτελέσματα (Σχήμα 3.2.1) όσο αφορά το evaluation set αρχικά για την Neuroshell2 και έπειτα στην Matlab (Σχήμα 3.2.2), όπου έδωσα περισσότερη έμφαση.

Neuroshell2 :

	Total	Training	Evaluation
Normal	444	377	67
Predicted		374	61
		98.41%	91.05%
Abnormal	239	202	37
Predicted		201	36
		98.02%	97.30%
Combined		99.31%	93.27%

Σχήμα 3.2.1 Καλύτερα αποτελέσματα Evaluation Set (Neuroshell2)

Matlab :

	Total	Training	Evaluation
Normal	444	377	67
Predicted		371	64
		98.41%	95.52%
Abnormal	239	202	37
Predicted		198	33
		98.02%	89.19%
Combined		98.27%	93.26%

Σχήμα 3.2.2 Καλύτερα αποτελέσματα στο Evaluation Set (Matlab)

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα με τρεις διαφορετικούς διαχωρισμούς training-evaluation set με τρεις προσπάθειες.

1 ^η Δοκιμή			2 ^η Δοκιμή			3 ^η Δοκιμή		
1	2	3	1	2	3	1	2	3
95,19%	93,27%	94,23%	98,07%	97,11%	95,19%	96,15%	95,19%	94,23%

Σχήμα 3.2.3 Διασταυρωμένη επικύρωση

Μερικά από τα καλύτερα αποτελέσματα για το συγκεκριμένο παράδειγμα σύμφωνα με άλλες μελέτες και με το συγκεκριμένο αρχείο εισόδου φαίνονται στον πιο κάτω πίνακα [13], [14].

Algorithm	Evaluation Set
Hyper Cube algorithm	98.2 %
RBF	96.18%
PNN	97%
MLP	95.74%
GRNN	98.8%

Σχήμα 3.2.4 Καλύτερα αποτελέσματα στο Evaluation Set από άλλες μελέτες

Δοκιμάστηκαν διάφορα άλλα παραδείγματα όπως για παράδειγμα το πρόβλημα του Pima Indians Diabetes, το οποίο έτρεξα στην Matlab με τα ακόλουθα καλύτερα αποτελέσματα :

	Total	Training	Evaluation
Normal	500	450	50
Predicted		404	39
		89.77%	78%
Diabetes	268	241	27
Predicted		141	19
		58.50%	70.37%
Combined		78.87%	75.32%

Σχήμα 3.2.5 Καλύτερα αποτελέσματα στο Evaluation Set

Μερικά από τα καλύτερα αποτελέσματα για το συγκεκριμένο παράδειγμα σύμφωνα με άλλες μελέτες και με το συγκεκριμένο αρχείο εισόδου φαίνονται στον πιο κάτω πίνακα [15].

Algorithm	Train	Test
MLP	81.08 %	77.60 %
RBF	100 %	68.23 %
GRNN	82.99 %	80.21 %

Σχήμα 3.2.4 Καλύτερα αποτελέσματα στο Evaluation Set από άλλες μελέτες

Μέσα από την διαδικασία αυτή, κατανόησα καλύτερα τόσο το λογισμικό Neuroshell2 αλλά και την Matlab, αλλά και επίσης μπόρεσα να μάθω πώς το κάθε εργαλείο πέρα από τις ρυθμίσεις παραμέτρων μπορεί να κάνει προ-επεξεργασία και μετά-επεξεργασία των δεδομένων κάτι το οποίο σε αρκετές περιπτώσεις μπορεί να δώσει καλύτερα αποτελέσματα.

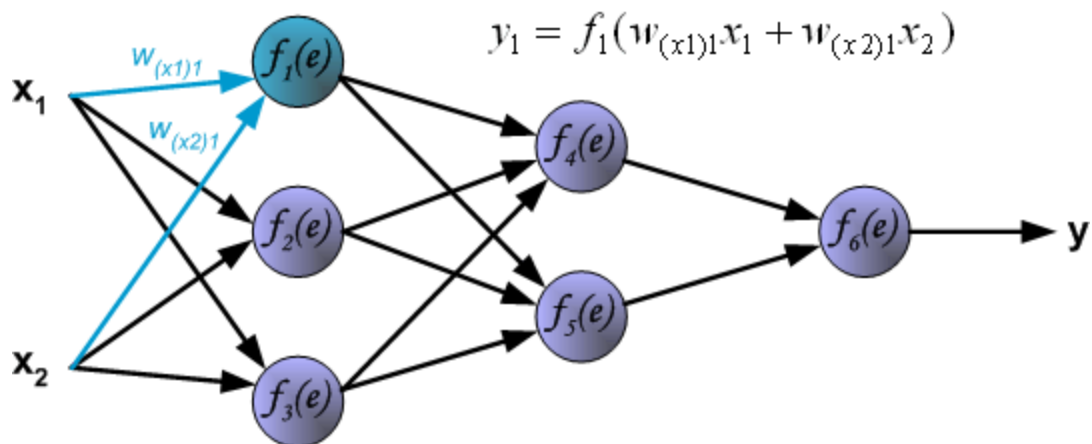
Ακόμη, χρησιμοποίησα τη μέθοδο διασταυρωμένης επικύρωσης, στην οποία ουσιαστικά παίρνουμε κάθε φορά διαφορετικό κομμάτι από το αρχείο με το οποίο εκπαιδεύουμε το δίκτυο. Έπειτα εκπαιδεύουμε το δίκτυο με τις ίδιες παραμέτρους πάνω στα διαφορετικά κομμάτια-δεδομένα εισόδου, έτσι ώστε να είμαστε βέβαιοι ότι το υπολογιστικό σύστημα όντως έχει μάθει και τα αποτελέσματά του είναι ορθά ανεξαρτήτως του αρχείου εισόδου.

3.3 Αλγόριθμοι – Τεχνικές που ακολουθήθηκαν

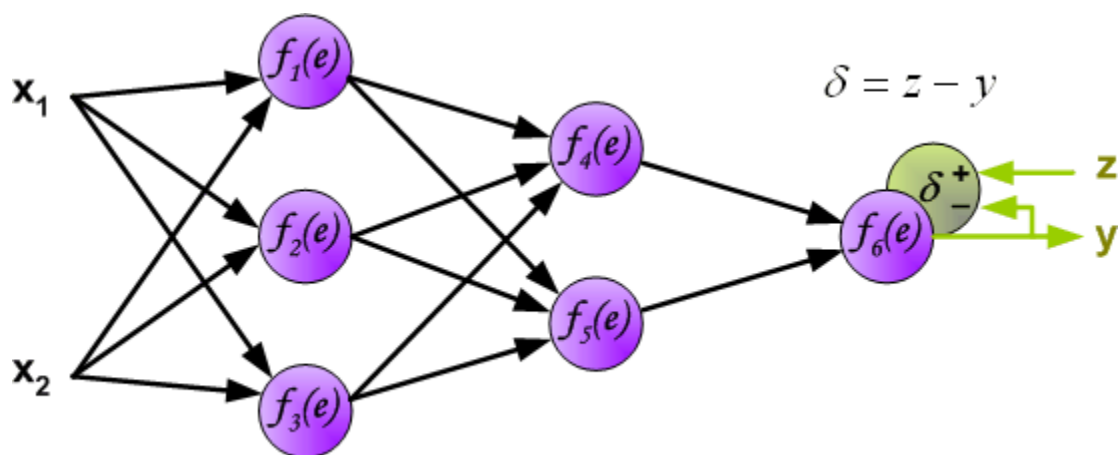
3.3.1 Αλγόριθμος πίσω διάδοσης λάθους

Ο αλγόριθμος που χρησιμοποίησα όσο αφορά τα νευρωνικά δίκτυα ήταν αυτός του Αλγόριθμου πίσω διάδοσης λάθους. Ο αλγόριθμος αυτός, ουσιαστικά χωρίζεται σε 2 φάσεις. Την διάδοση (Propagation) η οποία είναι και προς τα εμπρός (Διάδοση προς τα εμπρός της εισόδου μέσα από το δίκτυο έτσι ώστε να προκύψει η έξοδος) αλλά και προς τα πίσω (Διάδοση προς τα πίσω του αποτελέσματος μέσω του δικτύου χρησιμοποιώντας το επιθυμητό αποτέλεσμα για την εύρεση του σφάλματος και την διάδοση του σε όλο το δίκτυο) και επίσης την αλλαγή της τιμής των βαρών έτσι ώστε να μειωθεί το σφάλμα.

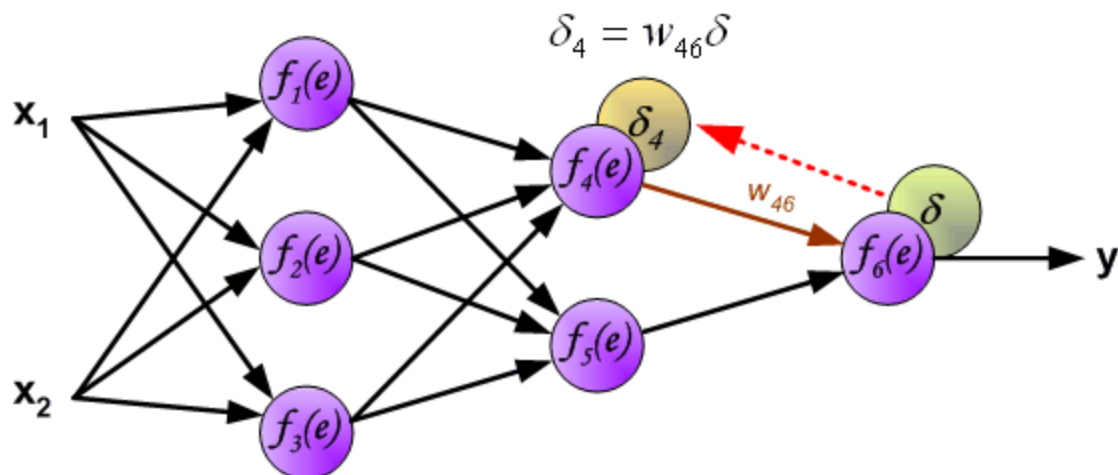
Οι εικόνες πιο κάτω δείχνουν την διάδοση του σήματος στο δίκτυο όπου το $w_{(xm)n}$ αντιπροσωπεύει τα βάρη της εισόδου x_m και του νευρώνα n στο επίπεδο εισόδου. Το y_n αντιπροσωπεύει το σήμα εξόδου του νευρώνα n .



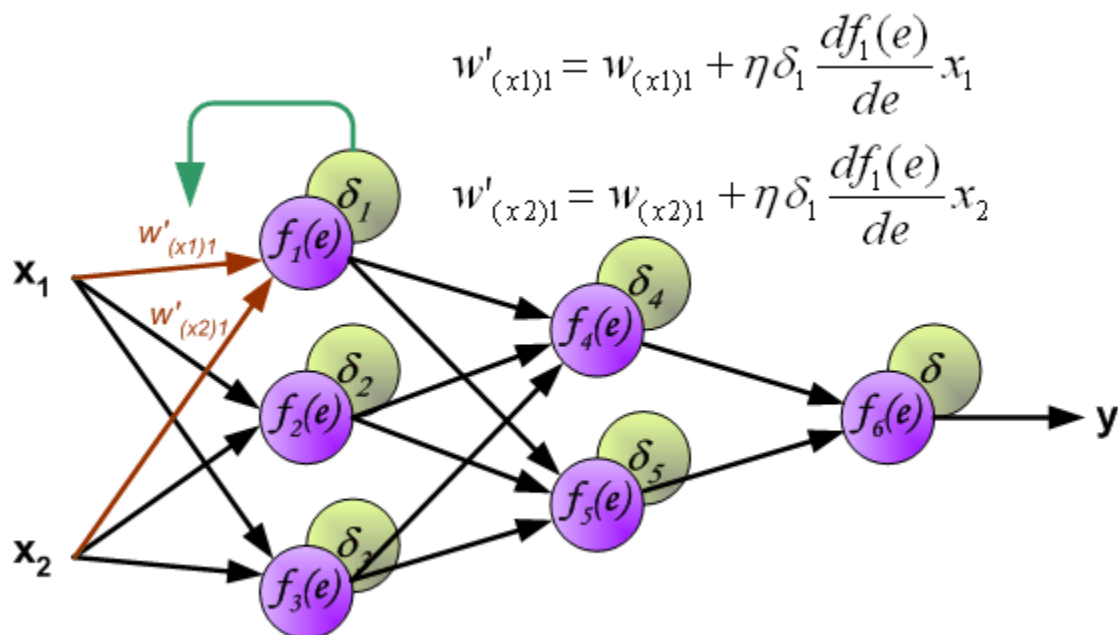
Εικόνα 3.1.1.1 Πολλαπλασιασμός της τιμής εισόδου με τα βάρη εισόδου και διάδοση μέσα από το δίκτυο μέχρι το επίπεδο εξόδου (y) για κάθε νευρώνα του δικτύου [16]



Εικόνα 3.1.1.2 Υπολογισμός του σφάλματος (δ) μεταξύ του αποτελέσματος που υπολογίσαμε και του επιθυμητού [16]



Εικόνα 3.1.1.3 Διάδοση του σφάλματος προς τα πίσω σε όλους τους νευρώνες του δικτύου και υπολογισμός τοπικού σφάλματος για κάθε νευρώνα [16]



Εικόνα 3.1.1.4 Υπολογισμός νέας τιμής βάρους για τον κάθε νευρώνα και διαδικασία υπολογισμού νέας εξόδου με ιδανικά μικρότερο σφάλμα [16]

3.3.2 Μηχανές Υποστήριξης Διανυσμάτων

Στην προσπάθεια να βρω καλύτερες λύσεις – μεγαλύτερο ποσοστό επιτυχίας για το πρόβλημα πρόβλεψης για το σύνδρομο Down με το ίδιο αρχείο περιστατικών, στράφηκα και στις μηχανές υποστήριξης διανυσμάτων.

Ουσιαστικά, χρησιμοποίησα κώδικα ο οποίος έπαιρνε ένα τυχαίο δείγμα από το αρχείο εισόδου για εκπαίδευση του δικτύου. Μετά από την διαδικασία χαρτογράφησης και διαχωρισμού των δεδομένων ακολουθούσε η αξιολόγηση του δικτύου. Αρχικά, άφησα τον αλγόριθμο να εκτελεί πολλές επαναλήψεις καταγράφοντας περιπτώσεις όπου τα ποσοστά επιτυχίας ήταν μεγαλύτερα από 90% έτσι ώστε να βρω το κατάλληλο εύρος των τιμών για τις παραμέτρους των μηχανών υποστήριξης διανυσμάτων που θα μου έδιναν τα καλύτερα αποτελέσματα.

Επίσης, σε μια προσπάθεια εξακρίβωσης των λόγων όπου τα ποσοστά επιτυχίας δεν ξεπερνούν κάποιο συγκεκριμένο ποσοστό (γύρω στο 96%), στον συγκεκριμένο αλγόριθμο καθόρισα ακολούθησα την εξής τακτική:

- Καθόρισα ένα κατώφλι ελάχιστου ποσοστού επιτυχίας 90%.
- Για οποιαδήποτε αρχιτεκτονική – παραμέτρους ικανοποιούσε αυτό το ποσοστό, αποθήκευα σε ένα πίνακα όλες τις περιπτώσεις για τις οποίες η πρόβλεψη ήταν λανθασμένη.
- Άφησα τον υπολογιστή να τρέχει για μερικές μέρες και συγκεκριμένα για 485 περιπτώσεις με τιμές sigma 1-5, penalty 2-100 και HalfSetSize=100.
- Ακολούθως ακολούθησα την ίδια διαδικασία αφήνοντας τις παραμέτρους στα ίδια όρια εκτός από την HalfSetSize (HalfSetSize=75), επειδή η παράμετρος αυτή δεν μπορεί να αλλάξει δυναμικά στο πρόγραμμα.
- Τέλος, μέτρησα την συχνότητα εμφάνισης της κάθε περίπτωσης μητέρας στο σύνολο των λανθασμένων προβλέψεων.

3.3.3 Υβριδικό Μοντέλο Γενετικού Αλγόριθμου – Νευρωνικών Δικτύων

Η προσέγγιση αυτή έχει σαν στόχο ουσιαστικά να βρει την καλύτερη αρχιτεκτονική και παραμέτρους για ένα νευρωνικό δίκτυο για οποιοδήποτε πρόβλημα κατηγοριοποίησης.

Υλοποίησα ένα γενετικό αλγόριθμο στην Matlab, του οποίου ο αρχικός πληθυσμός δημιουργείται τυχαία, με τρόπο τέτοιο ώστε οι δυαδικές τιμές που θα προκύψουν να αντιστοιχούν με τιμές μεταβλητών σε ένα νευρωνικό δίκτυο.

Χρησιμοποίησα διάφορες τιμές για εποχές και μέγεθος πληθυσμού και κατέληξα σε μέγεθος πληθυσμού 10 με 50 επαναλήψεις – εποχές όπου είχα αρκετά καλά αποτελέσματα, αλλά και αφού ο χρόνος εκτέλεσης του προγράμματος ήταν ήδη αρκετά μεγάλος.

Όσο αφορά κάποιες παραμέτρους για τον αλγόριθμο νευρωνικών δικτύων στο σημείο της αξιολόγησης, χρησιμοποίησα 100 επαναλήψεις – εποχές με συνάρτηση μάθησης για το νευρωνικό δίκτυο την συνάρτηση `trainingdm()` της Matlab, η οποία ανανεώνει τις τιμές των βαρών και του bias σύμφωνα με την μεθοδο κατάβασης κλίσης με ορμή. Σαν κανόνα αλλαγής των βαρών και του bias χρησιμοποίησα την συνάρτηση `'adaptwb'` η οποία πάντα ακολουθείται από την `'learnp'` (κανόνας αλλαγής Perceptron). Για υπολογισμό της απόδοσης του δικτύου, χρησιμοποίησα την `'mse'` (mean squared error) η οποία δεδομένων των πραγματικών αποτελεσμάτων, του επιθυμητού αποτελέσματος, καθώς και το σφάλμα στις τιμές των βαρών χρησιμοποιεί για υπολογισμό τον μέσο όρων των τετραγώνων των τιμών των σφαλμάτων [21].

Αναλυτικότερα σε κάθε βήμα το οποίο εκτελείται 50 φορές-εποχές για τα βήματα 2-9 :

- **Βήμα 1** : Συγκεκριμένα, στο 1^ο βήμα του γενετικού αλγόριθμου έχω κωδικοποιήσει σε δυαδική μορφή τις τιμές για τις παραμέτρους ρυθμού μάθησης, αριθμό νευρώνων κρυφού επιπέδου και συνάρτηση ενεργοποίησης για το κρυφό επίπεδο και το επίπεδο εξόδου όπου πιο κάτω φαίνονται το εύρος των τιμών.

- Ρυθμός μάθησης: 0.1 – 0.8
- Νευρώνες κρυφού επιπέδου: 1 – 512
- Συνάρτηση ενεργοποίησης κρυφού επιπέδου: { logsig, purelin, tansig, hardlim }
- Συνάρτηση ενεργοποίησης επιπέδου εξόδου: { logsig, purelin, tansig, hardlim }

Για τις τιμές του ρυθμού μάθησης, επειδή αλλάζει δυναμικά στο νευρωνικό δίκτυο και λόγω της υλοποίησης από την Matlab που δεν μπορεί να παραμείνει σε μια τιμή αλλά είναι μεταβαλλόμενη, διαιρούσα την μεταβλητή με το 1000, κάτι που είχε αρκετά καλά αποτελέσματα και έτσι αποφάσισα να κρατήσω και την κωδικοποίηση του learning rate.

- **Βήμα 2** : Ακολουθώς, αρχικοποιείται ο πληθυσμός μεγέθους 10 με τυχαίες τιμές για τις πιο πάνω μεταβλητές.
- **Βήμα 3 & 4** : Αφού γίνει η διαδικασία αποκωδικοποίησης για τις τιμές αυτές υπολογίζεται η αξιολόγηση κάθε χρωμοσώματος. Δηλαδή με τις τιμές που προκύπτουν, εκτελείται νευρωνικό δίκτυο με είσοδο τις δεκαδικές τιμές για τις πιο πάνω μεταβλητές και γίνεται η αξιολόγηση με βάση τον αριθμό των ορθών προβλέψεων στο evaluation set με κατώφλι κατηγοριοποίησης 0.5.
- **Βήμα 5** : Σε αυτό το βήμα, γίνεται η επιλογή των χρωμοσωμάτων με τα καλύτερα αποτελέσματα από το προηγούμενο βήμα (συνάρτηση αξιολόγησης). Όσο μεγαλύτερη είναι η τιμή της συνάρτησης αξιολόγησης, τόσο πιθανότερο είναι να επιλεχθούν αφού αυτό γίνεται με βάση ποσοτικής πιθανότητας, δηλαδή η τιμή της συνάρτησης διαιρείται με το άθροισμα όλων των τιμών για κάθε χρωμόσωμα για την συνάρτηση αξιολόγησης.
- **Βήμα 6 & 7** : Σύμφωνα με πιθανότητα 40%, γίνεται διασταύρωση 2 χρωμοσωμάτων με έλεγχο όμως να μην ξεπεράσει το 50%. Επιλέγεται ένα τυχαίο σημείο διασταύρωσης στο μήκος της δυαδικής αναπαράστασης, και από το σημείο εκείνο και μετά τα χρωμοσώματα των 2 αρχικών αντιστρέφονται μεταξύ

τους, δημιουργώντας έτσι ένα παιδί-καινούργιο χρωμόσωμα- με διαφορετικές τιμές.

- **Βήμα 8** : Στο σημείο αυτό επιλέγεται και πάλι ένα τυχαίο σημείο διασταύρωσης στο μήκος της δυαδικής αναπαράστασης με πιθανότητα 5%, όπου το ψηφίο σε αυτή την περίπτωση αντιστρέφεται, δηλαδή αλλάζει από 0 σε 1 ή αντίστροφα.

Επίσης, δοκίμασα και δεύτερη συνάρτηση αξιολόγησης η οποία να δίνει βάρος είτε σε περιπτώσεις normal (μη ανάπτυξης συνδρόμου Down) είτε στην αντίθετη κατηγορία.

Συγκεκριμένα, είναι η πιο κάτω εξίσωση, όπου δοκίμασα διάφορες τιμές για τα α και β με σκοπό να δώσω περισσότερη έμφαση σε κάποια κατηγορία (περιπτώσεις ανάπτυξης είτε όχι του συνδρόμου)

$$\text{Evaluation} = (\text{number_of_correct_predictions_abnormal}) * \alpha + (\text{number_of_correct_predictions_normal}) * \beta$$

Κεφάλαιο 4

Αποτελέσματα

4.1 Μηχανές Υποστήριξης Διανυσμάτων	39
4.1.1 Προσπάθεια εξακρίβωσης προβληματικών περιπτώσεων	41
4.2 Υβριδικό μοντέλο Γενετικού αλγόριθμου – Νευρωνικών δικτύων	44

4.1 Μηχανές Υποστήριξης Διανυσμάτων

Όσο αφορά τις τιμές των παραμέτρων για τις μηχανές υποστήριξης διανυσμάτων, κατέληξα στα πιο κάτω συμπεράσματα – εύρος των μεταβλητών :

- Οι τιμές του sigma ιδανικά είναι μικρότερες του 5 (αύξηση ανά 0.5), διότι αυτό καθορίζει το πλάτος της γραμμής που κατηγοριοποιεί τα δεδομένα μας. Οπότε στην περίπτωση που έχουμε πολύ μεγάλο sigma θα έχουμε ουσιαστικά και υπερμάθηση (over fitting), διότι αναγκάζουμε την γραμμή αυτή να κινείται πολύ κοντά στα δεδομένα ως αποτέλεσμα αν έρθει κάποιο καινούργιο – άγνωστο να μην ξέρουμε σε ποια κατηγορία θα συμπεριληφθεί, ενώ πολύ μικρό sigma δεν θα κατηγοριοποιεί ορθά τα δεδομένα μας και θα υπάρχει μεγάλο σφάλμα και μη καλή κατηγοριοποίηση.

- Το penalty είναι ο μέγιστος αριθμός των περιπτώσεων όπου δεν ζητάμε από τον αλγόριθμο να τις κατηγοριοποιήσει, αλλά να τις αγνοεί. Ένα ψηλό penalty έχει σαν αποτέλεσμα όχι και τόσο καλή γενίκευση, αλλά αντίθετα παρατηρείται υπερμάθηση, οπότε η επιλογή του κατάλληλου penalty size μπορεί να επέλθει μόνο μέσα από εκτενή πειραματισμό. Οι τιμές που κατέληξα για το penalty size είναι 2-100.

Για παράδειγμα πιο κάτω φαίνονται μερικά αποτελέσματα για HalfSetSize=75(Σχήμα 4.1.1) και HalfSetSize=100(Σχήμα 4.1.2) όπου είχα τα καλύτερα αποτελέσματα.

HalfSetSize	75		
Sigma	5	4	5
Penalty	12	12	62
Success Rate (Normal)	91.75%	92.61%	95.07%
Success Rate (Abnormal)	94.64%	94.24%	95.69%

Σχήμα 4.1.1 Αποτελέσματα από SVM

HalfSetSize	100		
Sigma	4	2.5	5
Penalty	22	2	12
Success Rate (Normal)	95.26%	95.79%	94.55%
Success Rate (Abnormal)	94.64%	94.64%	95.65%

Σχήμα 4.1.2 Περισσότερα αποτελέσματα από SVM

4.1.1 Προσπάθεια εξακρίβωσης προβληματικών περιπτώσεων

Στην προσέγγιση που ακολούθησα όπως περιγράφεται στο κεφάλαιο 3.3.2, μετά από πολλές επαναλήψεις του αλγορίθμου SVM και καταγραφή κάθε φορά των περιπτώσεων που αδυνατούσε να προβλέψει ορθά το σύστημα, προέκυψε η πιο κάτω παρατήρηση:

Σύνολο 23 περιπτώσεις όπου το βρέφος δεν ανέπτυξε το σύνδρομο Down, τόσο στις πρώτες 485 επαναλήψεις αλλά και στις δεύτερες 995 (με κάποια διαφορετικά χαρακτηριστικά στον αλγόριθμο όπως εξηγούνται στο κεφάλαιο 3.3.2), προβλέφθηκαν σε όλες τις περιπτώσεις λανθασμένα, ενώ σε καμία περίπτωση όπου το βρέφος ανέπτυξε το σύνδρομο Down η πρόβλεψη δεν ήταν αντίθετη, τουλάχιστον σε όλες τις περιπτώσεις.

Πιο κάτω (Πίνακας 4.1.1 και Πίνακας 4.1.2), για κάθε περίπτωση λανθασμένης πρόβλεψης, είτε εμφάνισης είτε όχι του συνδρόμου Down, έχω συμπεριλάβει τις 40 περιπτώσεις με το υψηλότερο ποσοστό όπου φαίνεται η ταυτότητα της μητέρας και ο αριθμός των εμφανίσεων ως μη ορθής πρόβλεψης.

Συγκεκριμένα, φαίνεται η ταυτότητα της μητέρας, πόσες φορές έχει προβλεφτεί λάθος η εμφάνιση ή όχι της νόσου τόσο στην πρώτη περίπτωση (485 επαναλήψεις) αλλά και στην δεύτερη (995 επαναλήψεις).

Στις τελευταίες δύο στήλες αναγράφεται το ποσοστό αποτυχημένης πρόβλεψης σύμφωνα με τον αριθμό των επαναλήψεων.

Ο πρώτος πίνακας αφορά περιπτώσεις False Negative, δηλαδή που είχαμε εμφάνιση συνδρόμου Down αλλά δεν προβλέφτηκε και στο δεύτερο πίνακα περιπτώσεις False Positive, δηλαδή μη εμφάνιση συνδρόμου Down αλλά πρόβλεψη ότι θα εμφανιζόταν.

ID	Εμφανίσεις 1(485)	Εμφανίσεις 2(995)	Ποσοστό 1	Ποσοστό 2
78824.1	469.000	941.000	97	95
93940.1	467.000	938.000	96	94
113161.1	408.000	820.000	84	82
48543.2	380.000	826.000	78	83
82404.2	362.000	746.000	75	75
83033.1	351.000	685.000	72	69
71156.1	328.000	692.000	68	70
86423.2	323.000	648.000	67	65
48288.2	322.000	625.000	66	63
91280.1	308.000	630.000	64	63
59243.1	294.000	626.000	61	63
98214.1	284.000	532.000	59	53
1131424.1	274.000	567.000	56	57
70333.1	253.000	518.000	52	52
113442.1	246.000	519.000	51	52
97256.1	234.000	487.000	48	49
79055.1	190.000	445.000	39	45
63660.1	189.000	383.000	39	38
72798.3	184.000	380.000	38	38
108542.1	183.000	405.000	38	41
72339.2	182.000	363.000	38	36
50613.2	178.000	391.000	37	39
77490.1	178.000	390.000	37	39
103722.1	176.000	389.000	36	39
100224.1	162.000	396.000	33	40
70639.1	154.000	355.000	32	36
81674.2	152.000	375.000	31	38
99510.1	146.000	315.000	30	32
63915.2	145.000	295.000	30	30
93882.1	140.000	335.000	29	34
34622.2	137.000	258.000	28	26
94099.1	134.000	280.000	28	28
58071.1	133.000	356.000	27	36
110035.1	126.000	290.000	26	29
85634.1	126.000	255.000	26	26
81849.1	125.000	298.000	26	30
58586.1	118.000	237.000	24	24
82005.1	113.000	261.000	23	26
71191.1	110.000	266.000	23	27

Πίνακας 4.1.1 Περίπτώσεις False Negative

ID	Εμφανίσεις 1(485)	Εμφανίσεις 2(995)	Ποσοστό 1	Ποσοστό 2
92525.1	485.0	995	100	100
78543.2	485.0	995	100	100
110964.1	485.0	995	100	100
108138.1	485.0	995	100	100
71165.1	485.0	995	100	100
102185.1	485.0	995	100	100
110710.1	485.0	995	100	100
105115.1	485.0	995	100	100
37328.2	485.0	995	100	100
59196.1	485.0	994	100	100
114257.1	485.0	994	100	100
140358.1	485.0	994	100	100
122295.1	485.0	994	100	100
69956.1	485.0	994	100	100
80900.1	485.0	993	100	100
120680.1	485.0	993	100	100
65456.1	485.0	992	100	100
59697.1	485.0	992	100	100
84818.2	485.0	992	100	100
82575.2	485.0	991	100	100
107029.1	485.0	991	100	100
73089.1	485.0	991	100	100
112918.1	484.0	995	100	100
101102.1	484.0	994	100	100
93753.1	484.0	993	100	100
111854.1	484.0	993	100	100
107508.1	484.0	993	100	100
83321.1	484.0	992	100	100
103980.1	484.0	992	100	100
73677.1	484.0	992	100	100
48511.4	484.0	992	100	100
10652.2	484.0	992	100	100
87471.1	485.0	990	100	99
114517.1	485.0	990	100	99
116386.1	485.0	989	100	99
103528.1	485.0	989	100	99
81217.1	485.0	988	100	99
58585.1	485.0	988	100	99
76708.2	485.0	984	100	99

Πίνακας 4.1.2 Περίπτώσεις False Positive

4.2 Υβριδικό μοντέλο Γενετικού αλγόριθμου – Νευρωνικών δικτύων

Στην προσέγγιση αυτή, χρησιμοποιήθηκε ο αλγόριθμος που περιγράφηκε στο Κεφάλαιο 3.3.3 και με λίγα λόγια είναι ένα μοντέλο γενετικού αλγόριθμου το οποίο χρησιμοποιεί σαν συνάρτηση αξιολόγησης το αποτέλεσμα του νευρωνικού δικτύου (τις ορθές προβλέψεις με κατώφλι 0.5).

Στο 1^ο βήμα του γενετικού αλγόριθμου

Αρχικά, χρησιμοποιήθηκε σαν συνάρτηση αξιολόγησης ο συνολικός αριθμός των ορθών προβλέψεων, όπου τα καλύτερα αποτελέσματα με αυτή τη συνάρτηση αξιολόγησης για κάθε εποχή του γενετικού αλγόριθμου φαίνονται στους πιο κάτω πίνακες :

Activation function : $\{0,1,2,3\} = \{ \text{logsig}, \text{purelin}, \text{tansig}, \text{hardlim} \}$

Ρυθμός Μάθησης : Ρυθμός Μάθησης /1000

Διαχωρισμός Δεδομένων:

	Normal Cases	Abnormal Cases
Training Set	25258	447
Evaluation Set	25259	246

Πίνακας 4.2.1 1^{ος} Διαχωρισμός Δεδομένων αρχείου

	Συνάρτηση Ενεργοποίησης 1	Συνάρτηση Ενεργοποίησης 2	Αριθμός Νευρώνων	Ρυθμός Μάθησης	Συνολικό Ποσοστό Επιτυχίας
1	1	2	234	6	95.4009017839639
2	2	1	54	1	98.9688296412468
3	0	2	434	2	99.0354832385807
4	2	1	54	5	75.4205057831798
5	3	1	182	5	77.7886688884532
6	3	2	238	4	65.5204861791806
7	3	1	316	4	68.3787492648500
8	2	1	182	5	77.0672417173103
9	2	1	202	6	74.4755930209763
10	2	1	178	5	83.1209566751617
11	1	1	458	6	71.2095667516173
12	1	1	246	4	71.1507547539698
13	2	1	308	1	69.8216036071359
14	1	1	438	5	74.4677514212900
15	0	1	180	4	90.2293667908253
16	0	1	180	5	95.2087825916487
17	0	1	166	5	92.2328955106842
18	2	1	180	5	84.6853558125858
19	0	1	180	4	98.6198784552049
20	0	1	436	5	89.8490492060380
21	3	1	436	4	58.2474024701039
22	0	1	436	5	95.4793177808273
23	0	1	502	2	70.7273083709077
24	0	1	52	1	75.7067241717310
25	3	1	242	2	15.4910801803568
26	0	1	488	5	89.4530484218781
27	3	1	242	2	93.7698490492060
28	3	1	242	2	45.4224661831014
29	3	1	296	7	68.1905508723780
30	0	1	232	5	95.3146441874142
31	1	1	424	7	61.9525583218977
32	0	1	204	5	97.3534601058616
33	1	1	480	3	89.3824740247010
34	3	1	232	5	75.6204665751813
35	1	1	232	5	69.4608900215644
36	3	1	416	3	55.8517937659283
37	3	1	200	5	87.4259949029602
38	3	1	416	3	73.5267594589296
39	3	1	200	5	49.3824740247010
40	3	1	496	3	65.4734365810625
41	2	1	508	6	74.4834346206626
42	2	1	508	6	82.4387375024505
43	2	1	252	6	85.7361301705548

44	1	1	504	7	81.1762399529504
45	3	1	376	6	61.2507351499706
46	2	1	254	6	81.0782199568712
47	0	1	506	7	94.0639090374436
48	0	1	120	5	96.8868849245246
49	0	1	506	7	87.9474612821015
50	0	1	506	7	94.9617722015291
51	0	1	506	7	95.0519505979220

Πίνακας 4.2.2 Καλύτερα Αποτελέσματα Evaluation Set

Δοκιμάζοντας ξανά τις τιμές των κωδικοποιημένων μεταβλητών για τις καλύτερες περιπτώσεις σε νέο νευρωνικό δίκτυο, κάποια αποτελέσματα προέκυψαν με αρκετή διαφορά, όπου πιο κάτω φαίνονται οι καλύτερες αποδόσεις.

Case 3	Evaluation	Case 19	Evaluation
Normal	25259	Normal	25259
Predicted	25228	Predicted	24509
	99.88%		97.03%
Abnormal	246	Abnormal	246
Predicted	14	Predicted	44
	5.69%		17.89%
Combined	98.96%	Combined	96.26%

Case 48	Evaluation
Normal	25259
Predicted	24469
	96.87%
Abnormal	246
Predicted	110
	44.72%
Combined	96.36%

Αν και τα ποσοστά επιτυχίας είναι πολύ υψηλά συνολικά, δεν είναι ιδανικά διότι ο αριθμός των ορθών προβλέψεων για τις περιπτώσεις όπου είχαμε ανάπτυξη συνδρόμου Down ήταν πολύ μικρός. Λόγω όμως της δυσαναλογίας των δύο περιπτώσεων στο αρχείο μας (ανάπτυξης και μή), το συνολικό ποσοστό επιτυχίας είναι πολύ μεγάλο λόγω της μεγάλης επιτυχίας πρόβλεψης για τις περιπτώσεις που δεν αναπτύχθηκε τελικά η πάθηση.

Έτσι, χρησιμοποίησα σαν συνάρτηση αξιολόγησης τον πιο κάτω τύπο για να δώσω περισσότερη βαρύτητα και στις περιπτώσεις ανάπτυξης συνδρόμου Down.

$$\text{Evaluation} = \text{correct_predictions_abnormal} * 0.8 + \text{correct_predictions_normal} * 0.2$$

Στην προσέγγιση αυτή, ο αλγόριθμος μπορούσε να προβλέψει ορθά μόνο τις περιπτώσεις ανάπτυξης συνδρόμου Down σχεδόν με μηδενικά ποσοστά πρόβλεψης για τις υπόλοιπες περιπτώσεις. Αυτό ίσως να οφείλεται λόγω στον πολύ μικρό συγκριτικά αριθμό των περιπτώσεων ανάπτυξης συνδρόμου Down (246 έναντι 24469).

Δοκιμάζοντας αρκετές διαφορετικές τιμές που να δίνουν έμφαση είτε στην μι είτε στην άλλη περίπτωση, κατέληξα στον πιο κάτω τύπο με τον οποίο είχα τα καλύτερα αποτελέσματα.

$$\text{Evaluation} = \text{correct_predictions_abnormal} * 0.2 + \text{correct_predictions_normal} * 0.8$$

Τα καλύτερα αποτελέσματα με αυτή τη συνάρτηση αξιολόγησης φαίνονται στον πιο κάτω πίνακα (Πίνακας 4.2.3).

	Συνάρτηση Ενεργοποίησης 1	Συνάρτηση Ενεργοποίησης 2	Αριθμός Νευρώνων	Ρυθμός Μάθησης	Συνολικό Ποσοστό Επιτυχίας
1	0	2	0	2	0.8000000000000000
2	0	2	114	1	0.764289827307791
3	0	2	114	1	0.8000000000000000
4	0	3	258	5	0.8000000000000000
5	0	2	256	2	0.659873627913998
6	1	0	22	1	0.8000000000000000
7	1	0	22	3	0.8000000000000000
8	1	0	22	1	0.8000000000000000
9	1	3	276	2	0.8000000000000000

10	1	3	276	2	0.8000000000000000
11	1	3	326	3	0.8000000000000000
12	1	0	262	1	0.8000000000000000
13	3	0	176	1	0.8000000000000000
14	0	0	22	4	0.8000000000000000
15	2	0	22	5	0.8000000000000000
16	3	0	22	7	0.8000000000000000
17	3	0	310	3	0.8000000000000000
18	1	0	48	1	0.8000000000000000
19	2	0	30	1	0.8000000000000000
20	3	0	148	0	0.8000000000000000
21	1	3	22	5	0.8000000000000000
22	1	0	148	1	0.8000000000000000
23	1	3	22	5	0.8000000000000000
24	1	0	18	5	0.8000000000000000
25	3	2	22	3	0.796771721389173
26	1	1	22	0	0.664145340451781
27	1	0	18	4	0.8000000000000000
28	1	0	18	4	0.8000000000000000
29	2	3	374	5	0.8000000000000000
30	3	3	150	7	0.8000000000000000
31	1	2	150	5	0.800015835939665
32	2	1	214	7	0.262307920834464
33	1	0	6	4	0.8000000000000000
34	3	0	438	7	0.8000000000000000
35	3	2	438	7	0.423661790677846
36	0	3	146	4	0.8000000000000000
37	0	3	146	4	0.8000000000000000
38	0	3	118	7	0.8000000000000000
39	0	0	270	4	0.8000000000000000
40	0	0	270	0	0.8000000000000000
41	2	2	176	7	0.446904572691952
42	0	0	244	7	0.8000000000000000
43	0	3	94	5	0.8000000000000000
44	0	0	240	5	0.8000000000000000
45	1	3	122	4	0.8000000000000000
46	0	0	240	7	0.8000000000000000
47	3	0	248	3	0.8000000000000000
48	3	3	120	0	0.8000000000000000
49	1	2	250	5	0.696017035866150
50	3	2	166	7	0.853735720697799
51	1	2	230	7	0.790737069649488

Πίνακας 4.2.3 Καλύτερα Αποτελέσματα Evaluation Set με 2^η συνάρτηση αξιολόγησης

Δοκιμάζοντας ξανά τις τιμές των κωδικοποιημένων μεταβλητών για τις καλύτερες περιπτώσεις σε νέο νευρωνικό δίκτυο, κάποια αποτελέσματα προέκυψαν με αρκετή διαφορά, όπου πιο κάτω φαίνεται η καλύτερη-πιο ισοζυγισμένη απόδοση.

	Evaluation
Normal	25259
Predicted	23391
	92.60%
Abnormal	246
Predicted	129
	52.43%
Combined	93.11%

Activation function 1: **tansig**

Activation function 2 : **hardlim**

Ρυθμός Μάθησης : **5/1000**

Αριθμός Νευρώνων : **22**

Κεφάλαιο 5

Συμπεράσματα

5.1 Συμπεράσματα διπλωματικής εργασίας

50

5.1 Συμπεράσματα διπλωματικής εργασίας

Η σημασία πρόβλεψης μιας ιατρικής πάθησης όπως για παράδειγμα αυτή του συνδρόμου Down είναι κάτι πολύ σημαντικό για μελλοντικούς γονείς και πρέπει να εκμεταλλευτούμε την συνεχή εξέλιξη της τεχνολογίας και την υπολογιστική ισχύ που έχουμε στα χέρια μας για να βρούμε λύσεις σε τέτοια σημαντικά προβλήματα.

Έχει αποδειχτεί πάμπολλες φορές ότι τα υπολογιστικά συστήματα και η τεχνητή νοημοσύνη είναι τα πλέον κατάλληλα για τέτοια προβλήματα και πολλά άλλα που αφορούν προβλήματα κατηγοριοποίησης.

Συγκεκριμένα, τα καλύτερα αποτελέσματα στην διπλωματική αυτή εργασία προέρχονται από τις μηχανές υποστήριξης διανυσμάτων με ποσοστό επιτυχούς πρόβλεψης για τις εντελώς άγνωστες περιπτώσεις κοντά στο 95%, με το υβριδικό μοντέλο γενετικού αλγόριθμου και νευρωνικών δικτύων να βρίσκεται πολύ κοντά στα ποσοστά αυτά με συνολικό ποσοστό επιτυχούς πρόβλεψης για τις εντελώς άγνωστες περιπτώσεις στο 93.11%.

Μέσα από την μέθοδο του υβριδικού μοντέλου που αναπτύχθηκε, τα αποτελέσματα είναι πάρα πολύ ενθαρρυντικά. Αυτό διότι τα ποσοστά επιτυχίας είναι πολύ κοντά σε σύγκριση με προηγούμενες προσπάθειες [29] και μάλιστα με σχεδόν διπλάσιο όγκο δεδομένων αλλά και πολύ μεγαλύτερο αριθμό εντελώς άγνωστων περιπτώσεων (Verification-Evaluation set). Επίσης λόγω της ευελιξίας που έχει το πρόγραμμα αλλάζοντας την συνάρτηση αξιολόγησης ή και με περαιτέρω πρόσθεση – αφαίρεση καινούργιων παραμέτρων τα ποσοστά αυτά δεν μπορούν παρά μόνο να βελτιωθούν.

Επίσης, κάτι που μελετήθηκε στο κεφάλαιο 4.1.1 που αφορά κάποιες επανειλημμένα λανθασμένες προβλέψεις, ίσως να οφείλεται και σε κάποια λάθη στις μετρήσεις του αρχείου που χρησιμοποιήσαμε λόγω του ανθρώπινου παράγοντα.

Κεφάλαιο 6

Προτάσεις για μελλοντική εργασία

6.1 Προτάσεις για μελλοντική εργασία

52

6.1 Προτάσεις για μελλοντική εργασία

Μετά από τα ενθαρρυντικά αποτελέσματα αυτής της διπλωματικής εργασίας, θα πρέπει να ακολουθηθούν περισσότερες προσεγγίσεις σαν αυτή που μελετήθηκε με τον υβριδικό αλγόριθμο Νευρωνικών δικτύων και Γενετικού αλγόριθμου. Αυτό διότι μέσα από υβριδικά συστήματα μπορούμε να συνδυάσουμε τα πλεονεκτήματα δύο προσεγγίσεων με αποτέλεσμα μια πιο αυτοματοποιημένη και επιστημονική λύση.

Μια πρόταση που αφορά τον συγκεκριμένο αλγόριθμο, είναι να κωδικοποιηθούν περισσότερες παράμετροι για το υβριδικό αυτό μοντέλο, όπως για παράδειγμα δύο κρυφά επίπεδα. Επίσης, μια άλλη παράμετρος που μπορεί να κωδικοποιηθεί είναι το πλάτος (c) της σιγμοειδούς συνάρτησης όπως φαίνεται στον τύπο $s_c(x) = 1/(1 + e^{-cx})$ [17]. Ψηλότερες τιμές για την μεταβλητή c , φέρνουν την συνάρτηση πιο κοντά σε αυτή της συνάρτησης κατωφλίου όπου ο νευρώνας τηρεί τον απλό κανόνα να πυροβολεί ή όχι. Με μικρότερες τιμές ωστόσο, η συνάρτηση τείνει να γίνεται περισσότερο γραμμική και να μην πυροβολεί. Αυτό ουσιαστικά μας επιτρέπει να κινούμαστε μεταξύ διαφορετικών ορίων δημιουργώντας μιας συνάρτησης η οποία πιθανόν να έχει καλύτερα αποτελέσματα από τις υπάρχουσες.

Μπορεί ακόμη να δοκιμαστούν διαφορετικοί αλγόριθμοι νευρωνικών δικτύων εκτός από τον αλγόριθμο πίσω διάδοσης λάθους, καθώς και εντελώς διαφορετικές αρχιτεκτονικές όπως για παράδειγμα αυτή του Hyper Cube Algorithm [14].

Μια τελευταία εισήγηση είναι η χρησιμοποίηση της τεχνολογίας επεξεργασίας εικόνας που χρησιμοποιείται πλέον ευρέως στην ιατρική [23]. Η προσέγγιση αυτή μπορεί να βοηθήσει στην ανάλυση ακτινογραφιών και συγκεκριμένα για εντοπισμό χαρακτηριστικών του βρέφους, που όπως αναφέρονται στο κεφάλαιο 2.1, πιθανό να οδηγήσουν σε εμφάνιση του συνδρόμου Down. Αυτό ενδεχομένως να οδηγήσει και σε αυτοματοποιημένη εξόρυξη δεδομένων μηδενίζοντας τις πιθανότητες καταγραφής λανθασμένων αποτελεσμάτων.

Βιβλιογραφία

- [1] Χρίστος Ν. Σχίζας, “Εισαγωγή στην Υπολογιστική Νοημοσύνη”, Συστήματα Υπολογιστικής Νοημοσύνης – Σημειώσεις Διαλέξεων, Τμήμα Πληροφορικής, Πανεπιστήμιο Κύπρου 2010.
- [2] McCulloch, Warren S., and Walter Pitts. „A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics*, 5 (1943): 115-133. Reprinted in *The Collected Works of Warren S. McCulloch*, vol. 1, Edited by Rook McCulloch, Salinas, CA: Intersystems Publications, 1989: 343-361.
- [3] <http://www.statsoft.com/textbook/support-vector-machines/>
- [4] Rosenblatt, Frank. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington, DC: Spartan Books, 1962.
- [5] <http://www.mind.ilstu.edu/curriculum/modOverview.php?modGUI=212>
- [6] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#What%20is%20a%20Neural%20Network
- [7] <http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=5069>
- [8] <http://www.fetalmedicine.com/fmf/>
- [9] <http://www.ncbi.nlm.nih.gov/pubmed/17973334>
- [10] <http://www.wardsystems.com/neuroshell2.asp>
- [11] Νεοφύτου Φοινίκη. “Σύστημα υπολογιστικής νοημοσύνης για διάγνωση στην ιατρική: Η περίπτωση του συνδρόμου Down”. Ατομική διπλωματική εργασία, Πανεπιστήμιο Κύπρου 2011
- [12] <http://archive.ics.uci.edu/ml/>
- [13] <http://www.yildiz.edu.tr/~tulay/publications/Tainn2003-5.pdf>
- [14] Dominic Palmer-Brown, Chrisina Jayne. Hypercube Neural Network Algorithm for Classification
- [15] Kamer Kayaer, Tulay Yildirim. Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks

- [16] http://home.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html
- [17] <http://page.mi.fu-berlin.de/rojas/neural/chapter/K7.pdf>
- [18] Costas K. Neocleous, Panagiotis Anastasopoulos, Kypros H. Nikolaides, Christos N. Schizas, Kleanthis C. Neokleous. Neural networks to estimate the risk for preeclampsia occurrence.
- [19] Elias Ebrahimzadeh, Mohammad Pooyan. Early detection of sudden cardiac death by using classical linear techniques and time-frequency methods on electrocardiogram signals.
- [20] Ghalia Ashoor, Nerea Maiz, Howard Cuckle, Firas Jawdat1 and Kypros H Nikolaides. Maternal thyroid function at 11–13 weeks of gestation in fetal trisomies 21 and 18.
- [21] <http://www.mathworks.com/help>
- [22] <http://www.wardsystems.com/manuals/neuroshell2>
- [23] <http://www.dsg.harvard.edu/uploads/Margarita/neuralnetworks011.pdf>
- [24] <http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001992/>
- [25] McCulloch, Warren S., and Walter Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity." Bulletin of Mathematical Biophysics, 5 (1943): 115-133. Reprinted in The Collected Works of Warren S. McCulloch, vol. 1, Edited by Rook McCulloch, Salinas, CA: Intersystems Publications, 1989: 343-361.
- [26] Corinna Cortes, Vladimir Vapnik: Support-Vector Networks. Machine Learning (ML) 20(3):273-297 (1995).
- [27] Χρίστος Χριστοδούλου, Υπολογιστικά Συστήματα Μάθησης – Σημειώσεις Διαλέξεων, Τμήμα Πληροφορικής, Πανεπιστήμιο Κύπρου 2010.
- [28] John Henry Holland. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence (1975).
- [29] Neocleous C.C., Nikolaides K.H., Neokleous K.C., Schizas C.N. (2010). First trimester diagnosis of Trisomy-21 by using artificial neural networks. Proceedings of the MEDICON 2010 Conference, Porto Carras, Greece, 2010

[30] J. Langdon H. Down, M.D., London. Observations on an Ethnic Classification of Idiots

[31] J. LEJEUNE, M. GAUTIER and R. TURPIN. Les chromosomes humains en culture de tissus. C. R. Acad. Sciences, January 26, 1959