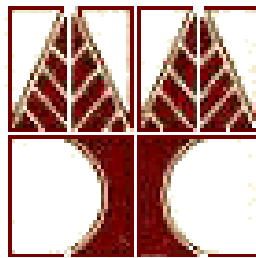


Ατομική Διπλωματική Εργασία

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

Νεοφύτου Φοινίκη

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2011

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Σύστημα Υπολογιστικής Νοημοσύνης για διάγνωση στην Ιατρική:
Η περίπτωση του συνδρόμου Down

Νεοφύτου Φοινίκη

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 2011

Ευχαριστίες

Θα ήθελα να εκφράσω θερμές ευχαριστίες στον επιβλέποντα Καθηγητή μου Χρίστο Σχίζα, που με εμπιστεύθηκε για την εκπόνηση της διπλωματικής εργασίας αυτής, όπως και τους βοηθούς του, Κλεάνθη Νεοκλέους και Αντρέα Νεοκλέους. Τους ευχαριστώ για την υπομονή και τη συνεχή καθοδήγησή τους καθώς και για την πολύτιμη βοήθειά τους.

Θα ήθελα να ευχαριστήσω επίσης και την οικογένειά μου για την ατέλειωτη υποστήριξη που μου προσφέρουν.

Περίληψη

Στόχος της διπλωματικής εργασίας αυτής ήταν η εύρεση ενός νευρωνικού δικτύου που θα ασχολείται με την σωστή διάγνωση και πρόβλεψη γενετικών ανωμαλιών, κατά τη διάρκεια μιας εγκυμοσύνης. Αποδεικνύουμε με αυτό τον τρόπο ότι τα νευρωνικά δίκτυα και γενικότερα ο κλάδος της τεχνητής νοημοσύνης μπορεί να φανεί πολύτιμος, ακόμη και στην ιατρική και την πρόοδό της.

Στο νευρωνικό σύστημα που κατασκευάστηκε προσπάθησα να αφομοιώσω ένα τεράστιο όγκο δεδομένων (34528 περιπτώσεις) από πραγματικά περιστατικά, από έγκυες γυναίκες, τα στοιχεία των οποίων πήραμε από το Fetal Medicine Foundation, στο Λονδίνο. Τα δεδομένα αυτά εφόσον πέρασαν από κάποια επεξεργασία, προσπάθησα να βρω την κατάλληλη αρχιτεκτονική έτσι ούτως ώστε το σύστημα να δίδει τα καλύτερα αποτελέσματα. Στη προσπάθεια αυτή με βοήθησε το λογισμικό Neuroshell2. Οι τρισωμίες που μελετήθηκαν είναι οι τρισωμίες 13(Patau syndrome), 18(Edward's syndrome) και 21(Down syndrome) καθώς και το Triploidy και το Turner Syndrome που είναι λίγο πιο σπάνιες γενετικές ανωμαλίες.

Επικεντρώθηκα και προσπάθησα να μελετήσω σε βάθος, τα feed forward δίκτυα με ένα, δύο ή τρία κρυφά επίπεδα. Τα αποτελέσματα στη αρχή ήταν μέτρια, αφού το σύστημα στη αρχή δεν έδειχνε να μαθαίνει λόγω της δυσαναλογίας στις περιπτώσεις. Είχαμε δηλαδή 511 περιπτώσεις με κάποια γενετική ανωμαλία και 34017 φυσιολογικές περιπτώσεις εγκυμοσύνης. Κάτι που όμως είναι φυσιολογικό, εφόσον τα περιστατικά αυτά είναι πιο σπάνια. Αυξάνοντας όμως την πολυπλοκότητα του δικτύου και τον αριθμό των κρυφών νευρώνων, και δοκιμάζοντας διάφορες συναρτήσεις ενεργοποίησης τα αποτελέσματα ήταν πολύ καλύτερα.

Τελικά το σύστημα έφτασε να δίνει πολύ καλά ποσοστά αποτελεσμάτων και να προβλέπει μέχρι και 76% των περιπτώσεων στο training set και μέχρι και 72,5% των περιπτώσεων στο evaluation set.

Ωστόσο, επιτυχία για ένα γιατρό για ένα τέτοιο σύστημα είναι σχετική και με βάση το τι έχει σήμερα και τι επιπλέον αποκτά με το προσφερόμενο σύστημα. Στη συγκεκριμένη περίπτωση τα αποτελέσματα είναι καλύτερα από αυτά που του δίνουν οι συμβατικές στατιστικές μέθοδοι που χρησιμοποιεί. Στόχος του γιατρού είναι να ελαχιστοποιήσει τις περιπτώσεις εγκύων γυναικών που θα χρειαστεί να κάνει διάγνωση κάποιας γενετικής παθολογίας όπως π.χ. τρισωμίας συνδρόμου Down με επεμβατικά μέσα (αμνιοκέντηση). Αυτό γιατί ένας γιατρός πρέπει να είναι σίγουρος προτού προβεί στην διάγνωσή του, αφού σε περιπτώσεις σαν αυτές παίζουμε με ανθρώπινες ζωές και δεν μπορούμε να οδηγηθούμε σε αποβολή του εμβρύου χωρίς να είμαστε απόλυτα σίγουροι. Το 76% και το 72,5%, είναι σαφώς πολύ καλά αποτελέσματα, αλλά δεν μπορούν να βοηθήσουν το γιατρό με απόλυτη σιγουριά, παρά μόνο σε συνδυασμό με την εμπειρία του, για να του αποκλείσουν κάποιες περιπτώσεις. Τα ποσοστά επιτυχίας μας όμως είναι μια καλή βάση που τέθηκε στο πρόβλημα αυτό, για να μπορεί να συνεχίσει η έρευνα και να φτάσουμε σε επίπεδα επιτυχίας της τάξης του 95% και άνω.

Περιεχόμενα

Κεφάλαιο 1	Θεωρία.....	1
1.1	Υπολογιστική Νοημοσύνη και Νευρωνικά δίκτυα	1
1.2	Τρισωμίες και Παράγοντες που επηρεάζουν	4
1.3	Patau Syndrome (Trisomy 13)	5
1.4	Edward's Syndrome (Trisomy 18)	6
1.5	Down Syndrome (Trisomy 21)	6
1.6	Triploid Syndrome	7
1.7	Turner Syndrome	8
1.8	Feed forward δίκτυα	8
1.9	Backpropagation αλγόριθμος	9
1.10	Προηγούμενες προσπάθειες για πρόβλεψη τρισωμιών	10
1.10.1	Προσπάθεια για πρόβλεψη της τρισωμίας 21	11
1.10.2	Προσπάθεια για πρόβλεψη της τρισωμίας 13,18, και 21	12
Κεφάλαιο 2	Ανάλυση Δεδομένων	14
2.1	Περιγραφή προβλήματος και μελέτη δεδομένων	14
2.2	Επεξεργασία δεδομένων	16
Κεφάλαιο 3	Μεθοδολογία που ακολουθήθηκε.....	19
3.1	Λογισμικά εργαλεία που χρησιμοποιήθηκαν	19
3.2	Μεθοδολογία	20
3.2.1	Από 35687 δεδομένα σε 34528	20
3.2.2	Προσπάθειες με χάρτη αυτοδιοργάνωσης Kohonen	20
3.2.3	Χωρισμός δεδομένων σε υποκατηγορίες	21
Κεφάλαιο 4	Περιγραφή λύσεων και αποτελέσματα.....	24
4.1	Αρχιτεκτονικές και αποτελέσματα	24
4.4.1	Αρχιτεκτονική με ένα κρυφό επίπεδο	26

4.1.2 Αρχιτεκτονική με δύο κρυφά επίπεδα	27
4.1.3 Αρχιτεκτονική με τρία κρυφά επίπεδα	28
4.2 Καλύτερα αποτελέσματα	32
Κεφάλαιο 5 Συμπεράσματα και μελλοντική εργασία	35
5.1 Συμπεράσματα	35
5.2 Μελλοντική εργασία	36
Βιβλιογραφία	38

Κεφάλαιο 1

Θεωρία

1.1 Υπολογιστική Νοημοσύνη και Νευρωνικά δίκτυα	1
1.2 Τρισωμίες και Παράγοντες που επηρεάζουν	4
1.3 Patau Syndrome (Trisomy 13)	5
1.4 Edward's Syndrome (Trisomy 18)	6
1.5 Down Syndrome (Trisomy 21)	6
1.6 Triploid Syndrome	7
1.7 Turner Syndrome	8
1.8 Feed forward δίκτυα	8
1.9 Backpropagation αλγόριθμος	9
1.10 Προηγούμενες προσπάθειες για πρόβλεψη τρισωμιών	10
1.10.1 Προσπάθεια για πρόβλεψη της τρισωμίας 21	11
1.10.2 Προσπάθεια για πρόβλεψη της τρισωμίας 13,18, και 21	12

1.1 Υπολογιστική Νοημοσύνη και Νευρωνικά δίκτυα

Υπολογιστική νοημοσύνη είναι κλάδος της επιστήμης της πληροφορικής, που ασχολείται με τη σχεδίαση και την υλοποίηση υπολογιστικών συστημάτων τα οποία, μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς. Τα στοιχεία αυτά μπορεί να είναι η μάθηση, η προσαρμοστικότητα, η εξαγωγή συμπερασμάτων, η κατανόηση από συμφραζόμενα, η επίλυση προβλημάτων.

Η κλασσική Τεχνητή Νοημοσύνη χρησιμοποιούσε συμβολική αναπαράσταση της γνώσης και τεχνικές ανάλυσης υψηλού επιπέδου (High-Level approach).

Η προσέγγιση αυτή είχε σοβαρούς περιορισμούς και έφτασε σε ένα σημείο που δεν μπορούσε να προσφέρει τίποτα καινούργιο. Η έρευνα στράφηκε σε τεχνικές ανάλυσης χαμηλού επιπέδου (Low-Level approach) .

Η βασική ιδέα ήταν η εξής: αντί να αναπαριστούμε την γνώση σε ένα υψηλό επίπεδο, αντικαθιστούμε τα μοντέλα αυτά με αλγορίθμους που στηρίζονται στην αριθμητική επεξεργασία της πληροφορίας (δηλαδή υπολογιστικά). Έτσι αντί να έχουμε ένα αναλυτικό μαθηματικό μοντέλο για την αγορά, κατασκευάζουμε έναν αλγόριθμο ο οποίος αναλύει και μαθαίνει από δεδομένα που ήδη υπάρχουν και περιγράφουν την αγορά.[6]

Κάποιες υποστηρικτικές μέθοδοι που χρησιμοποιούνται είναι:

1. Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)
2. Ασαφής Λογική (Fuzzy Logic)
3. Γενετικοί Αλγόριθμοι (Genetic Algorithms)

Μερικές εφαρμογές τεχνητής νοημοσύνης γενικότερα είναι:

- Λήψη αποφάσεων σε περιβάλλον Αβεβαιότητας
- Κόστος κύκλου ζωής
- Χρεολύσιο
- Τιμολόγηση αγαθών και υπηρεσιών
- Επιλογή προμηθευτή αγαθών
- Πρόβλεψη ζήτησης αγαθών και άλλων μεγεθών
- Αξιολόγηση διεργασιών και εξοπλισμού
- Χρονοπρογραμματισμός εργασιών και υπηρεσιών
- Προγραμματισμός παραγωγής
- Απόδοση επαίνου
- Αλγόριθμοι δρομολόγησης (fleet routing)

Τα νευρωνικά δίκτυα είναι το κομμάτι της υπολογιστικής νοημοσύνης με το οποίο ασχολήθηκαν για την εκπόνηση της εργασίας αυτής. Τα νευρωνικά δίκτυα, ορίζονται ως μαθηματικά μοντέλα για την επεξεργασία της πληροφορίας, η οποία υπολογίζεται μέσω συνάψεων, δηλαδή σημείων επαφής των νευρών. Τα νευρωνικά δίκτυα και η φιλοσοφία τους, είναι εμπνευσμένη από τα βιοηλεκτρικά δίκτυα που δημιουργούνται στον εγκέφαλο. Με πιο απλά λόγια, το νευρωνικό δίκτυο είναι ένα δίκτυο από υπολογιστικούς κόμβους (τους νευρώνες) που είναι συνδεδεμένοι μεταξύ τους έτσι όπως το κεντρικό νευρωνικό σύστημα, το οποίο προσπαθούν να αφομοιώσουν.

Εμπνεύστηκαν από τη βιολογία και έγιναν προσπάθειες μίμησης της ανοχής σε βλάβες και τη δυνατότητα εξόρυξης γνώσης μέσα από βιολογικά νευρωνικά συστήματα, μοντελοποιώντας τη δομή των χαμηλών επιπέδων του εγκεφάλου. Στη συνέχεια, από το κύριο μέρος της έρευνας της Τεχνητής Νοημοσύνης που έγινε το 1960-1980, προήλθαν τα Έμπειρα Συστήματα (Expert Systems). Τα συστήματα αυτά έχουν βασιστεί σε ένα μοντέλο λογικής διεργασίας υψηλού επιπέδου, το οποίο δημιουργήθηκε από τον τρόπο με τον οποίο έχει δομηθεί η συλλογιστική διεργασία των ανθρώπων σχετικά με το χειρισμό των συμβόλων. Έγινε γρήγορα εμφανές πως αυτά τα συστήματα αν και ήταν πολύ χρήσιμα σε κάποιους τομείς, απέτυχαν να συλλάβουν καίριες πτυχές της ανθρώπινης νοημοσύνης. Σύμφωνα με μια πτυχή της μελέτης, το γεγονός αυτό οφειλόταν στην αποτυχία τους να μιμηθούν τη βασική δομή του εγκεφάλου.

Προκειμένου να αναπαραχθεί νοημοσύνη, κρίνεται απαραίτητη η δημιουργία συστημάτων με παρόμοια τεχνική. Ο εγκέφαλος αποτελείται κατά κύριο λόγο από ένα ευρύ φάσμα νευρώνων (10.000.000.000 κατά προσέγγιση), οι οποίοι είναι μαζικά διασυνδεδεμένοι με ένα μέσο όρο από διάφορες χιλιάδες διασυνδέσεις ανά νευρώνα.

Κάθε νευρώνας είναι ένα εξειδικευμένο κύτταρο το οποίο έχει τη δυνατότητα μετάδοσης ενός ηλεκτροχημικού σήματος. Ο νευρώνας έχει μια διακλαδωτική διάρθρωση εισροών, τους δενδρίτες (dendrites), ένα κυτταρικό σώμα και μια διακλαδωτική δομή εκροών (τον άξονα). Οι άξονες ενός κυττάρου συνδέονται με τους δενδρίτες ενός άλλου, μέσω μιας σύναψης. Όταν, λοιπόν, ένας άξονας ενεργοποιηθεί, πυροδοτεί ένα ηλεκτροχημικό σήμα κατά μήκος του άξονα. Ένας νευρώνας εκτελεί αυτή τη διαδικασία μόνο όταν το συνολικό σήμα το οποίο λήφθηκε από τους δενδρίτες, υπερβεί ένα συγκεκριμένο επίπεδο, δηλαδή, το κατώτατο όριο βολής (firing threshold).

Η ισχύς ενός σήματος που λαμβάνεται από ένα νευρώνα, εξαρτάται από την αποτελεσματικότητα των συνάψεων. Κάθε σύναψη περιέχει ένα κενό με νευροδιαβιβαστές χημικών ουσιών (neurotransmitter chemicals) που είναι σε ετοιμότητα για μετάδοση ενός μηνύματος.

1.2 Τρισωμίες και παράγοντες που επηρεάζουν

Τρισωμίες στη γενετική έχουμε όταν τα έμβρυα παρουσιάζουν γενετικά λάθη , σχετικά με τον αριθμό των χρωματοσωμάτων στο γενετικό τους υλικό.

Ένας φυσιολογικός άνθρωπος γεννιέται με 46 χρωματοσώματα, 23 από τα οποία παίρνει από τον πατέρα και 23 από την μητέρα. Το κάθε χρωματόσωμα επηρεάζει και κάποια ανθρώπινα χαρακτηριστικά. Για παράδειγμα το 23^ο χρωματόσωμα που παίρνει το έμβρυο από τον πατέρα θα καθορίσει το φύλο του παιδιού. Η μητέρα πάντα δίνει X χρωματόσωμα 23. Αν ο πατέρας δώσει X τότε θα έχουμε XX άρα και κοριτσάκι, ενώ αν δώσει Y τότε θα έχουμε XY και επομένως το έμβρυο θα είναι αγοράκι.[1]

Σε ένα έμβρυο με τρισωμία, συμβαίνει η εξής γενετική διαταραχή. Αντί το παιδί να έχει 46 χρωματοσώματα, έχει 47. Άρα, αντί να έχει 2 χρωματοσώματα, ένα από τον κάθε γονιό, έχει σε κάποιο χρωματόσωμα 3 φορές. Επομένως ένα παιδί που έχει τρισωμία 21, έχει το χρωματόσωμα 21, τρεις φορές αντί για δύο. Το φαινόμενο αυτό ονομάζεται σύνδρομο Down (Down Syndrome). Ένα έμβρυο που έχει το χρωματόσωμα 13, τρεις φορές αντί για δύο, έχει Patau Syndrome, και αν έχει το χρωματόσωμα 18, τρεις φορές αντί για δύο, έχει Edward's Syndrome. Αυτά είναι τα τρία κύρια σύνδρομα που θα προσπαθήσω να προβλέπω, όσο το δυνατόν καλύτερα, μέσα από το νευρωνικό δίκτυο το οποίο θα κατασκευάσω.

Υπάρχουν βέβαια και άλλα δύο σύνδρομα με τα οποία θα ασχοληθούμε. Το triploid syndrome καθώς και το turner syndrome, που είναι λίγο πιο σπάνιες γενετικές ανωμαλίες, οι οποίες εξηγούνται πιο κάτω.

Προχωρώντας, σημαντικό είναι να βρούμε κάποιους παράγοντες, που σε κάποιο βαθμό ο καθένας, παίζουν ρόλο στις γενετικές αυτές διαταραχές.

1. Η ηλικία της μητέρας. Σύμφωνα με τα στατιστικά, η πιθανότητα το παιδί να παρουσιάσει σύνδρομο Down όταν η μητέρα είναι 20 ετών είναι 1:2000 γεννήσεις, ενώ όταν είναι 40 χρονών είναι 1:400 γεννήσεις.(MA)

2.CRL (Crown Rump length)

3.GA (Gestation Age) - η ηλικία της κύησης κατά τη μέτρηση.

4.Κληρονομικότητα από την οικογένεια ή ακόμα και αν η γυναίκα είχε προηγούμενη κύηση, με περιστατικό τρισωμίας.

5.NT (Nuchal Tranclucency)

6.FHR (Fetus Heard Rate) - παλμοί καρδιάς του εμβρύου

7.NB(Nasal Bone)-ρινικό κόκαλο

8.TF (Triscupid flow from RA to RV)-normal or not

9.DV(Ductus Venosus Flow)- normal or not- Ροή του φλεβώδους πόρου

10.Δείκτης του PAPP-A

11.B-hCG (Human Chorionic Gonadotropin)- χοριακή γοναδοτροπίνη[4][5]

1.3 Patau Syndrome (Trisomy 13)

Τρισωμία 13 έχουμε περίπου 1 στις 10000 γεννήσεις. Σημαντικό ρόλο εδώ, παίζει η ηλικία της μητέρας εφόσον ο μέσος όρος ηλικίας εντοπισμού τρισωμίας 13 είναι στα 31 χρόνια της μητέρας. Μερικά χαρακτηριστικά που μπορεί να παρουσιάσει το παιδί είναι τα ακόλουθα:

1.Δυσκολία στη σκέψη

- 2.Καρδιακά προβλήματα
- 3.Μη κανονικός παλμός
- 4.Χαμηλά αυτιά
- 5.Υποπλασία οπτικού νεύρου
- 6.Απλασία σε δέρμα και μαλλιά
- 7.Μικροφθαλμία
- 8.Επηρεασμοί στο συκώτι

1.4 Edward's Syndrome (Trisomy 18)

Η τρισωμία 18 είναι η δεύτερη πιο συχνή τρισωμία μετά το σύνδρομο down. Σημαντικό ρόλο στη πιθανότητα το παιδί να έχει αυτό το σύνδρομο, παίζει η ηλικία της μητέρας. Στατιστικά, Edward's syndrome, έχουμε 1 στις 3000 γεννήσεις.[2] Τα μωρά αυτά όμως συνήθως δεν καταφέρνουν να επιβιώσουν, και τα περισσότερα πεθαίνουν στη γέννα. Παρόλα αυτά αν καταφέρουν να επιβιώσουν, τα άτομα αυτά είναι πολύ πιο ευάλωτα σε ασθένειες παρά τα κανονικά άτομα. Μερικά χαρακτηριστικά που έχουν τα παιδιά αυτά είναι:

- 1.Δυσμορφίες στο συκώτι
- 2.Δομική ατέλεια στη καρδιά
- 3.Προβλήματα στο να φάει
- 4.Μικρό κεφάλι
- 5.Χαμηλά αυτιά και ασυνήθιστη μορφή
- 6.Υποπλασία κάτω γνάθου
- 7.Μύτη προς τα πάνω
- 8.Διασπασμένα χείλια
- 9.Στενές πτυχές βλεφάρων
- 10.Περισσότερη απόσταση από τον κανονικό ανάμεσα στα μάτια

1.5 Down Syndrome (Trisomy 21)

Το σύνδρομο down είναι μια από τις πιο συχνές τρισωμίες που παρουσιάζονται. Γι' αυτό και για κάθε εγκυμοσύνη γίνεται ένας προγεννητικός έλεγχος για να ελεγχθεί το έμβρυο και να προλάβουμε να προειδοποιήσουμε την μελλοντική μητέρα για πιθανόν

τέτοιου είδους εγκυμοσύνη. Μέχρι σήμερα οι παράγοντες που βοηθούσαν τους γιατρούς να καταλάβουν την τρισωμία 21 είναι[8]:

- 1.Μέτρηση της αυχενικής διαφάνειας πριν το τέλος του πρώτου τριμήνου
- 2.Αιματολογική ανάλυση από το μητρικό πλάσμα για την Α εμβρυακή πρωτεΐνη(AFP-A-Fetoprotein), για την Β χοριακή γοναδοτροπίνη (β-HCG) και για την ελεύθερη οιστραδιόλη(CUE3)
- 3.Προσδιορισμός καρυότυπου για λήψη χοριακών λαχνών
- 4.Ηλικία μητέρας
- 5.Κληρονομικότητα

Κάποια χαρακτηριστικά ενός παιδιού με down syndrome φαίνονται πιο κάτω:

- 1.Νοητική υστέρηση
- 2.Μη βραχυχρόνια μνήμη
- 3.Επηρεασμός σκέψης, κίνησης και ομιλίας
- 4.Καρδιαγγειακές βλάβες(π.χ. μεσοκοιλιακό έλλειμμα)
- 5.Ενδοκриноπάθειες (π.χ. διαβήτης)
- 6.Γλαύκωμα ή καταρράκτης

1.6 Triploid Syndrome

Τριπλοειδές σύνδρομο είναι το σύνδρομο κατά το οποίο το έμβρυο έχει 3 σειρές από χρωματοσώματα αντί για δύο που έπρεπε να έχει κανονικά. Όλα του τα χρωματοσώματα αντί να είναι δύο από το καθένα, ένα από τη μητέρα και ένα από τον πατέρα είναι τρία. Έτσι αντί για 46 χρωματοσώματα έχει 69. Το σύνδρομο αυτό είναι μια εξαιρετικά σπάνια γενετική διαταραχή η οποία αν προβλεφθεί έγκαιρα βοηθά τη μητέρα και το γιατρό να πράξουν καταλλήλως, ανάλογα με την περίπτωση. Τα μωρά που γεννιούνται με το σύνδρομο αυτό δεν ζουν συνήθως, για πάνω από 5 μήνες. Αυτά που καταφέρνουν να επιβιώσουν, έχουν μωσαϊσμό, δηλαδή κάποια κύτταρά τους έχουν κανονικό αριθμό

χρωματοσωμάτων (2) και κάποια άλλα είναι τριπλά[10]. Κάποια από τα συμπτώματα που μπορεί να φανούν στο έμβρυο είναι:

- 1.Μεγάλος πλακούντας
- 2.Ασύμμετρη ανάπτυξη του σκελετού
- 3.Φαρδιά μάτια
- 4.Χαμηλή μύτη
- 5.Χαμηλά αυτιά
- 6.Μικρή κάτω γνάθος

1.7 Turner Syndrome

Είναι το σύνδρομο [9] κατά το οποίο παρατηρείται διαγραφή ενός ολόκληρου χρωματοσώματος X. Είναι μια χρωμοσωμική αναταραχή στις γυναίκες, στην οποία το σύνολο ή μέρος ενός από τα χρωμοσώματα φύλου λείπει. Ένας φυσιολογικός άνθρωπος κανονικά, έχει 46 χρωμοσώματα, εκ των οποίων τα δυο είναι τα φυλετικά χρωμοσώματα. Οι γυναίκες έχουν XX και οι άντρες XY. Άρα στο σύνδρομο Turner, το ένα X από τα γυναικεία χρωμοσώματα ελλείπει ή έχει άλλες ανωμαλίες. Παρατηρείται περίπου σε 1 σε κάθε 2500 γεννήσεις. Κάποια από τα συμπτώματα που μπορεί να φανούν είναι:

- 1.Χαμηλό ανάστημα
- 2.Ευρύς θώρακας
- 3.Μικρότερο τέταρτο μετακάρπιο του χεριού
- 4.Διόγκωση χεριών και ποδιών
- 5.Πεταλοειδές νεφρό
- 6.Προβλήματα όρασης
- 7.Μικρότερα νύχια

1.8 Feed Forward δίκτυα

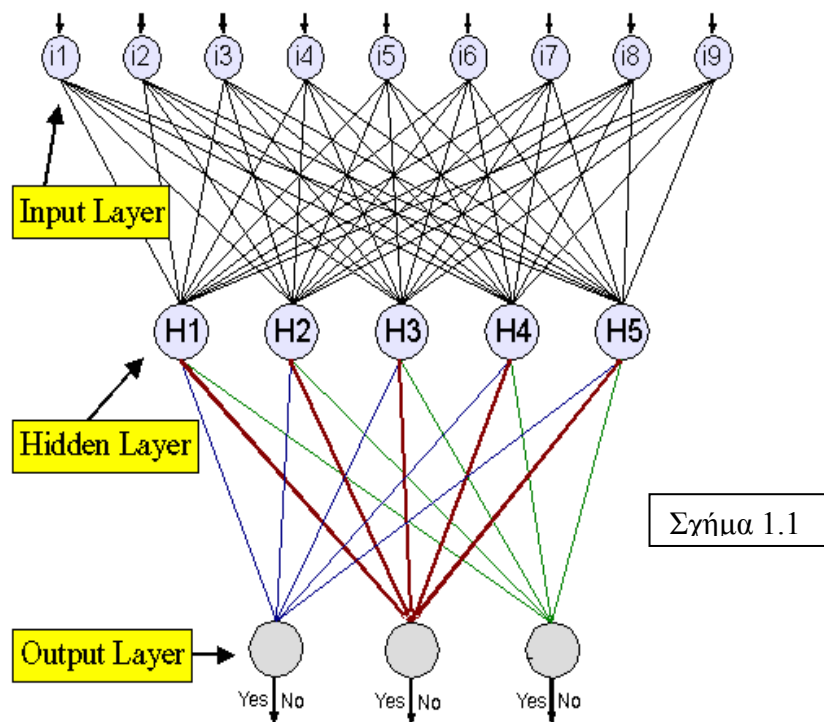
Οι λύσεις που θα καταγραφούν πιο κάτω είναι οι λύσεις των αρχιτεκτονικών που δοκιμάστηκαν για να καταλήξουμε στην καλύτερη λύση και τα συμπεράσματά μας. Οι δοκιμές, επικεντρώθηκαν σε αρχιτεκτονικές feed forward δικτύων, που χρησιμοποιούν back propagation αλγόριθμο, με εποπτεία.

Feed forward δίκτυα είναι τα δίκτυα όπου η πληροφορία ρέει από το επίπεδο εισόδου προς το επίπεδο εξόδου.

Στα δίκτυα με εποπτεία το σύνολο προτύπων περιέχει τόσο διανύσματα εισόδου όσο και διανύσματα εξόδου. Ο στόχος της εκπαίδευσης είναι η ελαχιστοποίηση του καθολικού σφάλματος και η προσαρμογή των τιμών των σωστών βαρών με τέτοιο τρόπο ώστε όταν διαβιβάζουμε στο δίκτυο τη σωστή είσοδο αυτό να αναπαράγει τη σωστή έξοδο.

1.9 Back propagation αλγόριθμος

Ο αλγόριθμος αυτός είναι ο ίσως ο πιο διαδεδομένος αλγόριθμος και βασίζεται στη επιβλεπόμενη μάθηση. Στην επιβλεπόμενη μάθηση έχουμε σαν δεδομένα από την αρχή τα δεδομένα εξόδου που θα πρέπει να βγάζει το σύστημα και ανάλογα με αυτά που πραγματικά μας βγάζει, υπολογίζουμε το λάθος του. Σκοπός είναι σε κάθε πέρασμα του αλγορίθμου(epoch) να μειώνουμε το σφάλμα του αυτό μέχρι να μπορεί να έλθει στο μικρότερο δυνατόν. Αυτό το πετυχαίνουμε αλλάζοντας τις τιμές των βαρών σε κάθε πέρασμα, εφόσον κάθε νευρώνας του δικτύου έχει και ένα βάρος από την αρχή.



Ο αλγόριθμος αυτός αποτελείται από τρεις φάσεις (Βλέπε σχήμα 2.1). Η πρώτη φάση είναι το εμπρόσθιο πέρασμα όπου αρχικοποιούνται τα βάρη και τα κατώφλια των διαφόρων συναρτήσεων ενεργοποίησης. Τα δεδομένα εισόδου δίνονται στο επίπεδο εισόδου, το οποίο τα δίνει στο κρυφό επίπεδο και προχωρεί προς τα μπροστά ανάλογα με το πόσα επίπεδα έχουμε. Η έξοδος του κάθε νευρώνα στο κρυφό επίπεδο (ή στα κρυφά επίπεδα αν είναι περισσότερα από ένα) καθώς και η έξοδος του κάθε νευρώνα στο επίπεδο εξόδου, υπολογίζεται από το άθροισμα των γινομένων των εισόδων τους. Το αποτέλεσμα τότε περνά από την ανάλογη συνάρτηση ενεργοποίησης (activation function) και στη συνέχεια στο επόμενο επίπεδο. Όταν φτάσουμε στο επίπεδο εξόδου μπορεί να υπολογιστεί τελικά το σφάλμα για το τρέχον πέρασμα. Τότε ξεκινά η δεύτερη φάση του αλγόριθμου back propagation, όπου παίρνουμε τις τιμές από τους νευρώνες του επιπέδου εξόδου και υπολογίζουμε το σφάλμα χρησιμοποιώντας τις επιθυμητές τιμές εξόδου. Έπειτα, όπως προϋποθέτει και το όνομα του αλγορίθμου, εφαρμόζουμε μια ανάστροφη μετάδοση του λάθους στο δίκτυο, ξεκινώντας από το επίπεδο εξόδου κινούμενοι προς τα πίσω.

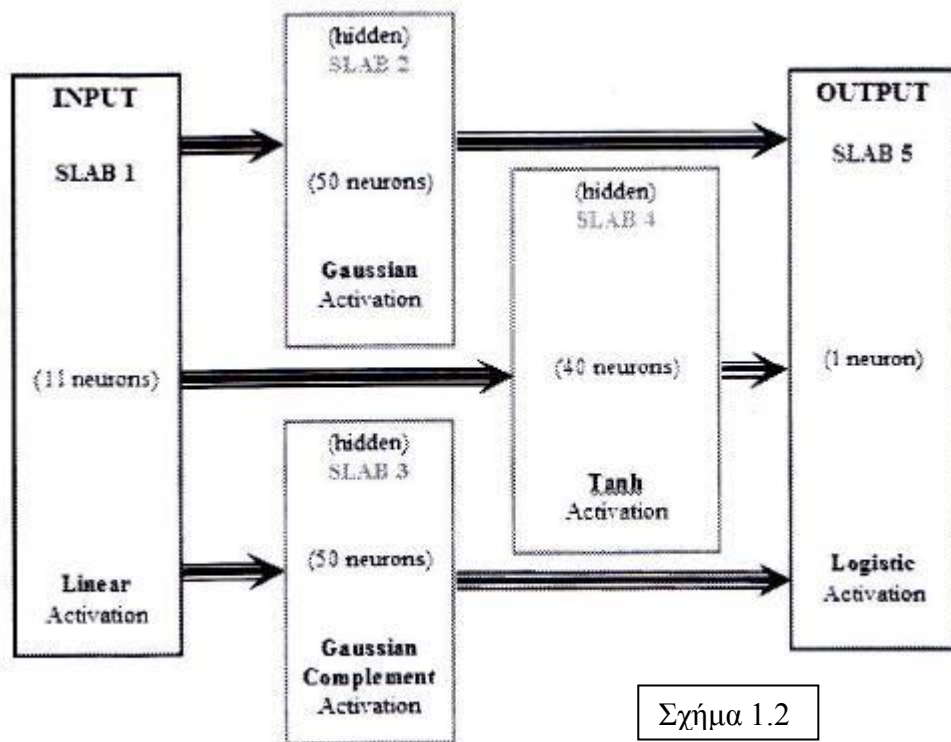
1.10 Προηγούμενες προσπάθειες για πρόβλεψη τρισωμίων

Όσον αφορά το Πανεπιστήμιο Κύπρου, στο παρελθόν έχουν ξαναγίνει προσπάθειες πρόβλεψης, για τις τρισωμίες σε συνδυασμό με τα νευρωνικά δίκτυα. Οι έρευνες έγιναν από τον Κ. Νεοκλέους, Κ. Νικολαΐδη, Κ. Νεοκλέους και Χ. Σχίζα και είναι δύο διαφορετικές.

Τα αποτελέσματα που εξάχθηκαν από τις έρευνες αυτές, ήταν λοιπόν πολύ ενθαρρυντικά για τη συνέχεια της έρευνας και για στη δική μας περίπτωση. Οι έρευνες αυτές έπαιξαν πολύ σημαντικό ρόλο και στην προσπάθεια εκπόνησης της διπλωματικής μου εργασίας, αφού μου έδωσαν κάποιο υπόβαθρο γνώσεων για το θέμα στο οποίο μπορούσα να στηριχτώ.

1.10.1 Προσπάθεια για πρόβλεψη της τρισωμίας 21

Το πείραμα αυτό διεξάχθηκε με 23513 περιπτώσεις ασθενών από το Fetal Medicine Foundation στο Λονδίνο. Από τα δεδομένα αυτά, μόλις στις 217 περιπτώσεις τα έμβρυα έπασχαν από τρισωμία 21, ενώ οι υπόλοιπες 23296 ήταν φυσιολογικές. Εδώ με την χρήση των νευρωνικών δικτύων, προσπάθησαν να προβλέψουν αν το έμβρυο θα έπασχε από σύνδρομο down, από μετρήσεις που έγιναν κατά το πρώτο τρίμηνο της



εγκυμοσύνης. Τα αποτελέσματα ήταν αρκετά καλά χρησιμοποιώντας μια multi-slab αρχιτεκτονική (Σχήμα 1.2) με 11 εισόδους που φαίνονται πιο κάτω (Πίνακας 1.1) και μία έξοδο η οποία έβγαζε 0, όταν προέβλεπε ότι το έμβρυο ήταν φυσιολογικό και 1 αν έπασχε από τρισωμία 21. Τα καλύτερα αποτελέσματα της αρχιτεκτονικής αυτής στο training set ήταν 98,9%, στο test set 100%, ενώ στο evaluation set έφτασαν μέχρι το 93,3%. Οι συναρτήσεις ενεργοποίησης (activation functions), καθώς και η πολυπλοκότητα των νευρώνων σε κάθε επίπεδο φαίνεται στο Σχήμα 1.2. [4]

Table 1 Parameters used for the first trimester diagnosis of trisomy 21

Maternal age
CRL, Crown Rump Length (mm)
GA, Gestation Age when the crown rump length (CRL) was measured (in days)
Previous T21, T18, T13 (yes or no)
NT, Nuchal Translucency (mm)
FHR, Fetus Heart Rate
NB, Nasal Bone (normal, abnormal)
TF, Tricuspid Flow from RA to RV (normal, abnormal)
DV, Ductus Venosus Flow (normal, abnormal)
Defects
Serum marker PAPP-A
β-hCG, Human Chorionic Gonadotropin

Πίνακας 1.1

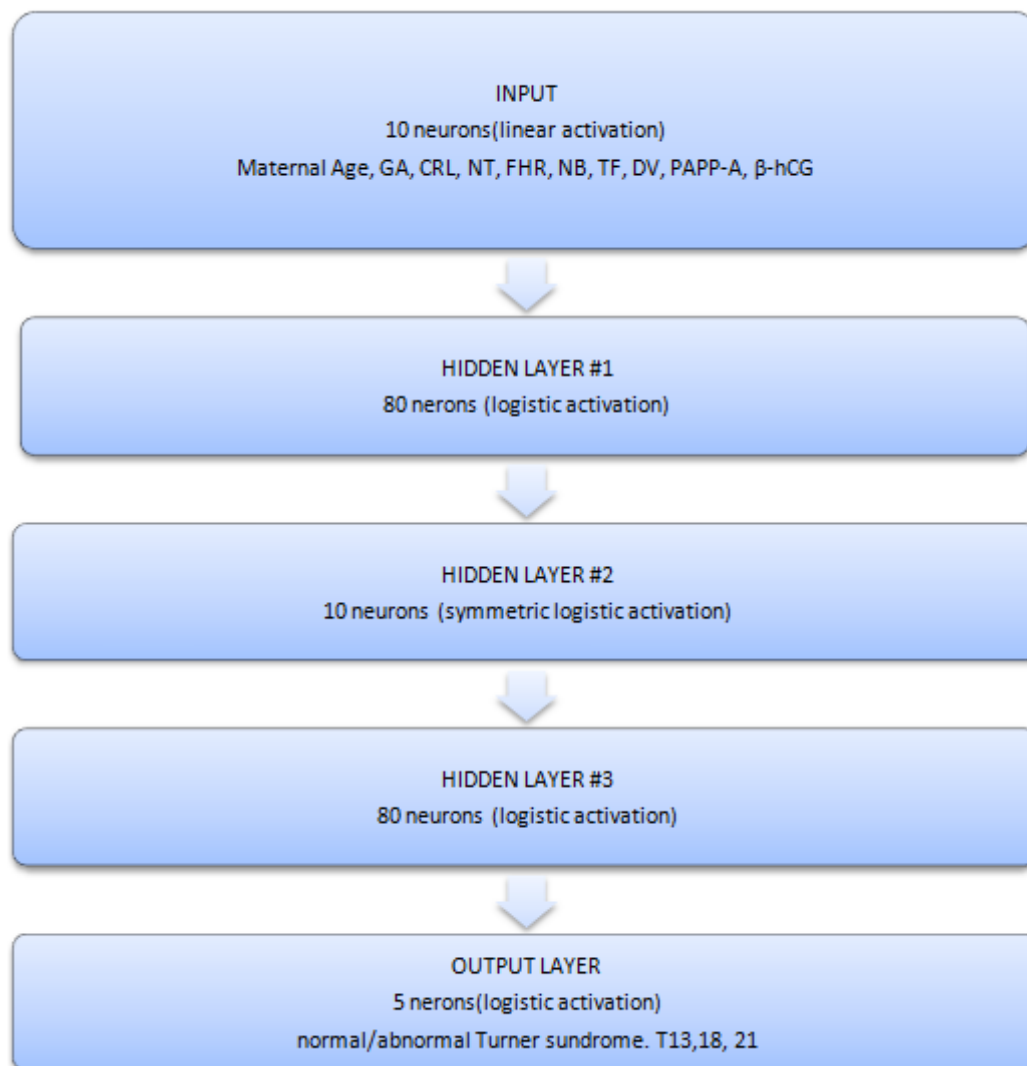
1.10.2 Προσπάθεια για πρόβλεψη της τρισωμίας 13,18, και 21

Η δεύτερη προσπάθεια, έγινε με διαφορετικά δεδομένα τα οποία περιείχαν, εκτός από τις φυσιολογικές περιπτώσεις ασθενών και τις περιπτώσεις με τρισωμία 21(σύνδρομο down), και άλλες περιπτώσεις τρισωμιών. Την τρισωμία 13, 18 και το Turner syndrome. Τα δεδομένα αυτά ήταν αυτή τη φορά 31611 περιπτώσεις από τις οποίες οι 31135 ήταν φυσιολογικές και οι 476 με κάποιο από τα σύνδρομα που μελετούσαν. Το δίκτυο που έβγαλε τα καλύτερα αποτελέσματα αυτή τη φορά ήταν ένα feedforward δίκτυο με 10 εισόδους και 3 κρυφά επίπεδα. Οι νευρώνες εξόδου ήταν 5, δηλαδή ένας για κάθε σύνδρομο και ένας που δήλωνε αν ήταν φυσιολογική περίπτωση εγκυμοσύνης ή με κάποια γενετική ανωμαλία γενικά. Οι είσοδοι που χρησιμοποιήθηκαν φαίνονται στον Πίνακα 1.2 πιο κάτω ενώ αρχιτεκτονική με τις συναρτήσεις ενεργοποίησης και την πολυπλοκότητα των νευρώνων στο Σχήμα 1.3.

- 1 Maternal age
- 2 GA Gestation age when the crown rump length was measured(days)
- 3 CRL, Crown rump length
- 4 NT Nuchal translucency
- 5 FHR Fetus heart rate
- 6 NB Nasal bone
- 7 TF Tricuspid flow from RA to RV (normal, abnormal)
- 8 DV Ductus venosus flow(normal, abnormal)
- 9 Serum maker PAPP-A
- 10 β-Hcg Human chorionic gonadotropin

Πίνακας 1.2

Τα εντελώς άγνωστα για το σύστημα δεδομένα του evaluation set, έφτασαν μέχρι και το ποσοστό 98,1%. Το σύστημα είχε επίσης ποσοστό επιτυχίας 85,1% στις άγνωστες για αυτό περιπτώσεις των εγκυμοσυνών που τα έμβρυα έπασχαν από κάποια γενετική ανωμαλία. Οι άγνωστες περιπτώσεις της τρισωμίας 21 προβλέφθηκαν κατά 78.9%, της τρισωμίας 18 κατά 79,2% , της τρισωμίας 13 κατά 0,0% ενώ του Turner syndrome κατά 42,9%. [5]



Σχήμα 1.3

Κεφάλαιο 2

Ανάλυση Δεδομένων

2.1 Περιγραφή προβλήματος και μελέτη δεδομένων	14
2.2 Επεξεργασία δεδομένων	16

2.1 Περιγραφή προβλήματος και μελέτη δεδομένων

Το πρόβλημα το οποίο η διπλωματική μου έχει ως στόχο λοιπόν, είναι η κατασκευή ενός συστήματος το οποίο θα συνδυάζει κάποιες μεθοδολογίες στατιστικών προσεγγίσεων μαζί με νευρωνικά δίκτυα για να μπορεί να κάνει μια σωστή διάγνωση για το αν το έμβρυο μιας εγκύου θα πάσχει από κάποια γενετική ανωμαλία ή όχι. Το σύστημα θα μπορεί να δέχεται σαν είσοδο κάποια δεδομένα από μετρήσεις διαφόρων παραγόντων που παίζουν κάποιο ρόλο και θα πρέπει να είναι σε θέση να κάνει κάποια πρόβλεψη χρησιμοποιώντας την εμπειρία για την οποία έχει εκπαιδευτεί.

Για τη διπλωματική μου, μου δόθηκαν κάποια δεδομένα προς μελέτη, τα οποία δόθηκαν από το νοσοκομείο Fetal Medicine Foundation, στο Λονδίνο, και είναι βασισμένα σε πραγματικά περιστατικά. Είναι βασικά κάποιες μετρήσεις από εξετάσεις που έγιναν σε έγκυες γυναίκες και βλέπουμε το αποτέλεσμα στο τέλος, αν δηλαδή το παιδί έχει κάποιο γενετικό πρόβλημα και ποίο είναι αυτό, όπως επίσης και ποια ήταν τα συμπτώματά του στην αρχή (δηλαδή τα αποτελέσματα από τις εξετάσεις).

Τα δεδομένα αυτά, ήταν αυτά που βασίστηκα από την αρχή για να εκπαιδεύσω το σύστημά μου. Βοήθησαν αρχικά στην έρευνα, για να δούμε ποιοι παράγοντες επηρεάζουν και μετριούνται από γιατρούς καθώς και πόσο επηρεάζουν. Έκανα δηλαδή, μια μελέτη με βάση τα δεδομένα και από το διαδίκτυο και από ανάλογα βιβλία για να γνωρίσω τις ασθένειες, να μάθω τι ήταν το καθετί που μετριόταν και πόσο αυτό επηρέαζε μια εγκυμοσύνη καθώς και πόσο συχνά επηρέαζε την έγκυο να παρουσιάσει

γενετικές ανωμαλίες. Στη συνέχεια κάποια από τα δεδομένα αυτά χρησιμοποιήθηκαν σαν είσοδοι στο σύστημά μας, και κάποια έμειναν τελείως άγνωστα προς το σύστημα για να μπορέσω μετά να ελέγχω και την απόδοσή του σε τελείως άγνωστες περιπτώσεις. Τα δεδομένα αυτά λοιπόν σαφώς, έπαιξαν καθοριστικό ρόλο στην εκπαίδευση του συστήματός.

Πιο κάτω λοιπόν περιγράφεται τι είναι η καθεμία από τις εννέα εισόδους που χρησιμοποιούμε στα δεδομένα και ποια η σημασία της:

MA-Maternal Age: σε αυτήν τη στήλη δεδομένων αποθηκεύεται η ηλικία της μητέρας αφού όπως πιο πάνω είδαμε, παίζει πολύ σημαντικό ρόλο στην εμφάνιση κάποιας γενετικής ανωμαλίας στο παιδί. Όσο πιο μεγάλη είναι μια γυναίκα, τόσες πιο πολλές πιθανότητες έχει.

GA-Gestation Age: είναι η ηλικία της κύησης, σε μέρες, δηλαδή σημειώνεται πόσων ημερών είναι ακριβώς το έμβρυο. Είναι επίσης σημαντική μέτρηση, διότι σε κάποια ηλικία κάποια πράγματα μπορούν να φανούν και κάποια όχι.

CRL-Crown Rump Length: μήκος του κεφαλουραίου.

NT-Nuchal Translucency: αυχενική διαφάνεια

PAPP-A: μια ουσία που μετρούν οι γιατροί για να αποφανθούν αν είναι σε φυσιολογικά πλαίσια ή όχι.

B-Hcg: είναι η χοριακή γοναδοτροπίνη, μια ουσία που μετρούν οι γιατροί για να αποφανθούν αν είναι σε φυσιολογικά πλαίσια ή όχι.

Prev: σαν είσοδος μπαίνει και το αν υπήρξε στη συγκεκριμένη γυναίκα περιστατικό, όπου στο έμβρυο εντοπίστηκε οποιαδήποτε μορφή τρισωμίας. Υπάρχει στήλη για την τρισωμία 21, για την 18 και την 13.

2.2 Επεξεργασία δεδομένων

Ανάλογα με το σύστημα, τροποποιήθηκαν βέβαια κατάλληλα και τα δεδομένα, χωρίς σαφώς να αλλάζει η πληροφορία τους. Για παράδειγμα, δεν μπορεί να γίνει στρογγυλοποίηση μιας μέτρησης γιατί ίσως χαλάσει το αποτέλεσμα, αλλά όταν τα δεδομένα μου εκφράζονται με yes/no ή με normal/abnormal, μπορώ να χρησιμοποιήσω 0 και 1. Έτσι γενικά βρέθηκαν κάποιοι τρόποι να κωδικοποιούνται τα δεδομένα όπου επιτρεπόταν. Ένα screenshot από τα δεδομένα φαίνεται πιο κάτω:

	ID Sort 34528 cases	GA-d	GA-w	MA	Prev T21Num (271)	Prev T18Num (105)	Prev T13Num (59)
1							
2	111688,1	88	12,5	34,9	0	0	0
3	111599,1	82	11,8	36,1	0	0	0
4	111513,1	82	11,7	38,2	0	0	0
5	111460,1	91	13,0	33,6	0	0	0
6	111352,1	89	12,7	34,9	0	0	0
7	111348,1	85	12,1	30,4	0	0	0
8	111114,1	95	13,5	36,2	0	0	0
9	111110,1	84	12,0	28,9	0	0	0
10	111105,1	90	12,9	29,9	0	0	0
11	111093,1	88	12,6	34,1	0	0	0
12	111076,1	95	13,6	29,8	0	0	0
13	111057,1	89	12,7	32,8	0	0	0

CRL	NT	PAPP-A MoM (733 NT)	b-hCG MoM (733 NT)	FHR (43 NT)	DefSimpleNum	NBNum (169 NT)	TrNum (135 NT)	DVNum (266 NT)	OutNum, 1-5	OutNum, 1, 0	T21	T18	T13	Triploidy	Turner	USE
61,1	4,7	0,667	0,564	168	0	1	0	0	0	0	0	0	0	0	0	T
51,2	6,1	1,047	1,821	146	0	1	1	1	0	0	0	0	0	0	0	V
50,3	6,9	0,235	0,340	165	1	1	0	1	2	1	Other GD	1	Other GD	Other GD	Other GD	T
67,0	3,5	1,143	1,577	151	0	0	0	0	0	0	0	0	0	0	0	V
62,8	2,4	0,752	2,480	162	0	0	1	1	1	1	1	Other GD	Other GD	Other GD	Other GD	T
55,6	2,2	0,727	0,739	153	1	0	0	1	0	0	0	0	0	0	0	T
74,9	2,2	0,828	9,066	157	0	0	0	0	0	0	0	0	0	0	0	T
54,3	4,7	1,219	1,510	172	0	0	0	0	0	0	0	0	0	0	0	V
66,0	3,6	0,418	0,567	158	0	0	0	0	0	0	0	0	0	0	0	T
61,7	2,4	1,475	3,094	165	0	1	0	0	0	0	0	0	0	0	0	T
75,4	3,6	0,900	3,631	153	0	0	1	0	1	1	1	Other GD	Other GD	Other GD	Other GD	T
63,3	5,6	0,976	0,855	154	0	0	0	0	0	0	0	0	0	0	0	T

Πίνακας 2.1

Όπως φαίνεται πιο πάνω οι στήλες με τα δεδομένα μας είναι 24 (Βλέπε πίνακα 2.1).

Στην πρώτη στήλη υπάρχει η ταυτότητα του ασθενή και στις υπόλοιπες μετρήσεις που έγιναν από εξετάσεις, κωδικοποιημένες ή μη. Στις στήλες prev21, prev18, prev13 έχουμε 0 εάν η έγκυος δεν αντιμετώπισε περιστατικό εγκυμοσύνης με προηγούμενη τρισωμία και 1 εάν έχει αντιμετωπίσει 21,18 ή 13 ανάλογα. Στη στήλη defects simple num έχουμε κωδικοποιημένα κάποια «ελαττώματα» που παρουσιάζει το έμβρυο:

0 Normal

1 Exomphalos

2 Exomphalos, Megacystis

3 Megacystis

4 Holoprocencephaly

5 Holoprocencephaly, Exomphalos

Στη στήλη output num(1-5) έχουμε αντίστοιχα ποια γενετική ανωμαλία έχει το έμβρυο τελικά μετά τη γέννησή του:

0 Normal

1 Trisomy 21

2 Trisomy 18

3 Trisomy 13

4 Triploidy syndrome

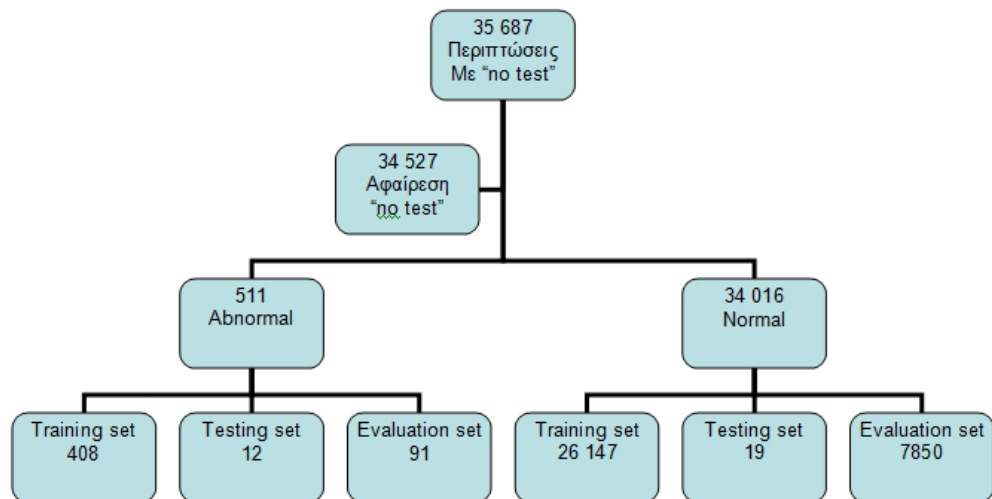
5 Turner syndrome

Στις στήλες NBNum, TriNum, DVNum, Output num, T21, T18, T13, Triploidy, και Turner έχουμε 0 για normal περιπτώσεις και 1 για abnormal περιπτώσεις. Στις υπόλοιπες στήλες οι μετρήσεις, όπως είναι φυσικό δεν μπορούσαν να κωδικοποιηθούν αλλά ούτε και να στρογγυλοποιηθούν, εφόσον θα χάναμε σημαντική πληροφορία με κίνδυνο να αλλοιωθεί το τελικό αποτέλεσμα. Τέλος, στη τελευταία στήλη, το USE, χωρίστηκαν τα δεδομένα σε 3 ομάδες αναλόγως του σκοπού που θα χρησιμοποιηθούν. Ποια θα ανήκουν στο training, ποια στο testing, και ποια στο evaluation set. T, P, και V αντίστοιχα. Κωδικοποιήθηκαν με τα γράμματα που αναγνωρίζει η Neuroshell2.

Επίσης τα δεδομένα που είχα ήταν αρχικά 35687 περιπτώσεις. Το πρόβλημα ήταν ότι μερικά από αυτά ήταν ελλιπή (δεν είχε καθόλου δεδομένα-no test) και έπρεπε να αφαιρεθούν τελείως για να μην επηρεάσουν το τελικό αποτέλεσμα.

Στη συνέχεια αφού τα δεδομένα μας κωδικοποιήθηκαν και αφαιρέθηκαν όσα δεν ήταν χρήσιμα, ήταν έτοιμα για να χρησιμοποιηθούν από αρχιτεκτονικές για να βγουν κάποια αποτελέσματα για αξιολόγηση. Κατά την κωδικοποίηση, με την στήλη USE όπως είδαμε πιο πάνω, χωρίστηκαν σε 3 διαφορετικά set. Το training, το testing και το evaluation. Ο διαχωρισμός έγινε με μια αναλογία στα κανονικά και τα προβληματικά

δεδομένα, έτσι ούτως ώστε να μην έχουμε προβλήματα εκμάθησης στο σύστημα. Ο διαχωρισμός φαίνεται πως έγινε στο σχήμα 2.1 που παρουσιάζεται πιο κάτω:



Σχήμα 2.1

Κεφάλαιο 3

Μεθοδολογία που ακολουθήθηκε

3.1 Λογισμικά εργαλεία που χρησιμοποιήθηκαν	19
3.2 Μεθοδολογία	20
3.2.1 Από 35687 δεδομένα σε 34528	20
3.2.2 Προσπάθειες με χάρτη αυτοδιοργάνωσης Kohonen	20
3.2.3 Χωρισμός δεδομένων σε υποκατηγορίες	21

3.1 Λογισμικά εργαλεία που χρησιμοποιήθηκαν

Τα δεδομένα, αυτό που θέλαμε ήταν να μπαίνουν σαν είσοδοι σε ένα νευρωνικό δίκτυο και να βγάζουν σαν αποτέλεσμα: 0 (normal) αν δεν προβλέπεται το έμβρυο να πάσχει από οποιαδήποτε ανωμαλία και 1 (abnormal) αν προβλέπεται ότι το έμβρυο πάσχει από κάποια γενετική ανωμαλία.

Για να πάρουμε όμως τα καλύτερα αποτελέσματα, έπρεπε να βρεθεί ένας τρόπος για να δοκιμαστούν αρκετές αρχιτεκτονικές χωρίς να χρειάζεται να σχεδιάζω και να προγραμματίζω κάθε τους κομμάτι από την αρχή. Έτσι χρησιμοποιήθηκαν κάποια έτοιμα λογισμικά που με βοήθησαν στην έρευνά μου.

Πριν τρέξω λοιπόν τα δεδομένα αυτά, και αποφασίσω ποιες αρχιτεκτονικές πρέπει να χρησιμοποιήσω για όσο το δυνατόν καλύτερα αποτελέσματα, έπρεπε να εξοικειωθώ με τα συστήματα αυτά. Μπήκα λοιπόν στα help τους, διάβασα κάποια πράγματα και έτρεξα κάποια παραδείγματα που με βοήθησαν. Τα συστήματα που χρησιμοποίησα ήταν το NeuroShell Classifier, το NeuroShell Predictor, και το NeuroShell 2. Ανάλογα με το ποιο θα μπορούσε να με βοηθήσει καλύτερα, αυτό θα χρησιμοποιούσα και πιο πολύ. Μια καλή ιδέα όμως ήταν να χρησιμοποιήσω ένα συνδυασμό, με βάση τις λειτουργίες που μπορεί να μου προσφέρει το καθένα.

Αφού δοκίμασα και πειραματίστηκα και με τα τρία συστήματα, κατάλαβα ότι περισσότερο σε αυτό που θέλω να κάνω, θα με βοηθούσε το NeuroShell 2. Αυτό γιατί στο NeuroShell 2 μπορείς να αλλάξεις διάφορες αρχιτεκτονικές και να παίζεις με πολλές διαφορετικές παραμέτρους του συστήματος.

Για να δοκιμάσω καλά το NeuroShell 2, πήρα κάποια παραδείγματα δεδομένων και από τα άλλα προγράμματα και τα είσαξα στο πρόγραμμα ώστε να τα επεξεργαστώ.

Δοκίμασα πολλές αρχιτεκτονικές και άλλαξα συνέχεια πολλές παραμέτρους μέσα σε αυτές. Έτσι δημιούργησα ένα πίνακα σ' ένα excel file, όπου κάθε φορά που άλλαζα οποιαδήποτε παράμετρο και έτρεχα ξανά τα δεδομένα μου, κατέγραφα τα αποτελέσματα. Με αυτό τον τρόπο μελέτησα αργότερα, με ποια αρχιτεκτονική και ποιες παραμέτρους το κάθε πρόβλημα λύθηκε, με τα πιο καλά αποτελέσματα.

3.2 Μεθοδολογία

3.2.1 Από 35687 δεδομένα σε 34528

Φτάνοντας λοιπόν και στα πραγματικά δεδομένα, ακολουθήθηκε η ίδια διαδικασία. Τα δεδομένα όμως επειδή ήταν χιλιάδες, χρειάστηκαν μια περεταίρω επεξεργασία πριν προχωρήσω. Εφαρμόστηκε αρχικά μια μείωση. Η μείωση αυτή έγινε σβήνοντας κάποια δεδομένα που δεν μας παρείχαν καμία πληροφορία. Δηλαδή σε μερικές από τις στήλες του αρχείου μας αντί το αποτέλεσμα της ανάλογης εξέτασης στην οποία υποβλήθηκε η έγκυος, αναγραφόταν η έκφραση “no test”. Αυτό σημαίνει ότι η έγκυος δεν είχε πραγματοποιήσει την εξέταση αυτή ή δεν υπήρχαν επαρκή στοιχεία. Επομένως όλα αυτά τα δεδομένα έπρεπε να φύγουν από το αρχείο για να μην επηρεάσουν το τελικό αποτέλεσμα. Εφόσον τα έλεγα και είδα ότι δεν θα μπορούσε να επηρεάσει σε κάτι η απουσία τους, διαγράφηκαν.

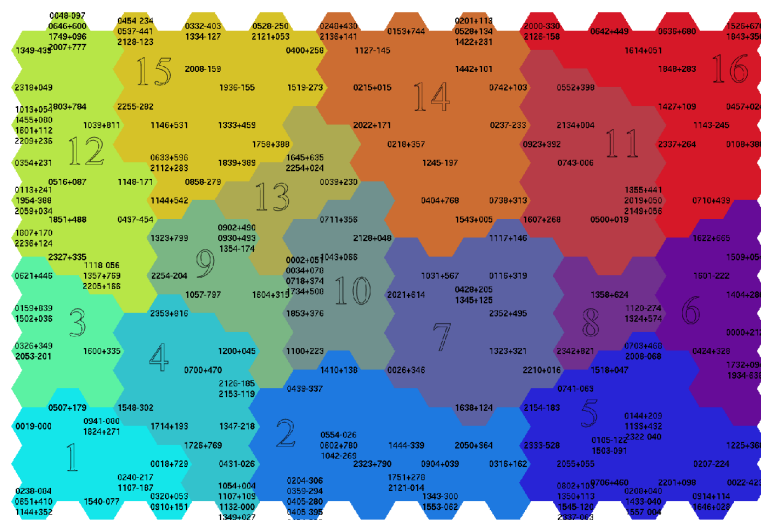
3.2.2 Προσπάθειες με χάρτη αυτοδιοργάνωσης Kohonen

Πριν προχωρήσω πιο κάτω, προσπάθησα να μειώσω περισσότερο τα δεδομένα μου για να βοηθήσω το σύστημά μου να μάθει πιο εύκολα κατά την διαδικασία της

εκπαίδευσής του. Προσπάθησα δηλαδή να βρω δεδομένα τα οποία είχαν τις ίδιες εγγραφές περίπου από τα normal data που ήταν χιλιάδες και να διαγράψω ένα ανάλογο ποσοστό, χωρίς να επηρεάζω το τελικό αποτέλεσμα. Δοκίμασα σε αυτήν την ομάδα δεδομένων λοιπόν να τρέξω τον αλγόριθμο του Kohonen, ο οποίος προσφέρει αυτοδιοργάνωση του χάρτη. Δηλαδή τα δεδομένα οργανώνονται μόνα τους σε ομάδες, μέσα στις οποίες βρίσκουν τα περισσότερα κοινά χαρακτηριστικά με άλλες εγγραφές της ομάδας. Η κάθε εγγραφή όμως δικαιούται να ανήκει μόνο σε μία ομάδα, οπότε οι εγγραφές όλες αναγκάζονται να επιλέξουν μια μόνο ομάδα με την οποία έχουν τα περισσότερα κοινά χαρακτηριστικά. Ο τρόπος που αυτοδιοργανώνεται ο χάρτης Kohonen φαίνεται στο παράδειγμα στο σχήμα 3.1.

Σχήμα 3.1

Clusters by centroid method, N = 16



Το πείραμά μας αυτό δεν πέτυχε στο τέλος διότι ο όγκος δεδομένων ήταν τεράστιος, και τα δεδομένα δεν χωρίστηκαν ακριβώς στις ομάδες όπως αναμέναμε. Δεν μπορούσα να διαγράψω λοιπόν τίποτα, αφού ίσως κάπου επηρεάζα το τελικό αποτέλεσμα. Αποφασίσαμε λοιπόν να προχωρήσουμε πιο κάτω με τα δεδομένα που έχουμε, για να δούμε πιο μπορεί να είναι τελικά το αποτέλεσμα με αυτά.

3.2.3 Χωρισμός δεδομένων σε υποκατηγορίες:

Αυτό που έπρεπε να γίνει πιο κάτω ήταν να χωριστούν τα δεδομένα ανάμεσα σε ποια από αυτά θα χρησιμοποιούνταν για training, ποια για testing, και ποια για evaluation set. Αυτό έγινε κάνοντας εισαγωγή μιας επιπλέον στήλης στο τέλος των data μου με το όνομα USE. Η διαδικασία μετά γίνεται αυτόματα μέσω της NeuroShell 2, χωρίζοντας τα δεδομένα όπως πρέπει σε ανάλογα αρχεία. Έτσι χρησιμοποιήθηκαν 26555 περιπτώσεις για training, 31 περιπτώσεις για testing και 7941 περιπτώσεις για evaluation. Οι αναλογίες των normal και abnormal περιπτώσεων φρόντισα να είναι σε όλα τα set δεδομένων οι ίδιες περίπου, έτσι ούτως ώστε, να μην έχουμε προβλήματα στη μάθηση του συστήματος. Φρόντισα επίσης να υπάρχουν και κατάλληλες αναλογίες περιπτώσεων σε όλα τα set δεδομένων, από όλες τις ασθένειες.

Το training, testing, και evaluation set, είναι ένας διαχωρισμός που πρέπει να γίνεται στα δεδομένα μας, όπως εξήγησα πιο πάνω, διότι η NeuroShell 2, χρησιμοποιεί αυτές τις 3 ομάδες δεδομένων για ένα διαφορετικό σκοπό, στη προσπάθειά της να εκπαιδεύσει το σύστημα μας:

Τα δεδομένα του **training set**, χρησιμοποιούνται στο να εκπαιδευτεί το σύστημά μας πάνω σε αυτά και έτσι τα δεδομένα του είναι γνωστά.

Τα δεδομένα που περιλαμβάνονται στο **testing set** είναι άγνωστα για το σύστημα αλλά τα χρησιμοποιεί για να βλέπει πόσο έμαθε το σύστημά μας μετά το τέλος κάθε εποχής. Βασικά, γίνεται κάποιος έλεγχος της προόδου του training. Αν τα αποτελέσματα σε κάποιο χρονικό σημείο t ήταν καλύτερα από ότι στο χρονικό σημείο $t-1$, τότε τα βάρη φυλάσσονται ως αυτά του καλύτερου σετ. Στο τέλος τα καλύτερα βάρη φυλάγονται για να χρησιμοποιηθούν κατά την πρόβλεψη.

Τα δεδομένα του **evaluation set** με τη σειρά τους, είναι τελείως άγνωστα για το σύστημα και τα τρέχουμε εμείς, αφού τελειώσει η εκπαίδευση ολόκληρου του συστήματος, για να δούμε αν η πρόβλεψη θα είναι εξίσου καλή και σε άγνωστα δεδομένα.[3]

Μετά από την πιο πάνω διαδικασία τα δεδομένα ήταν έτοιμα για εισαγωγή στο σύστημα. Σημείωσα λοιπόν κάτω κάποιες αρχιτεκτονικές, οι οποίες η μια μετά την

άλλη είχαν μια λογική μεταξύ τους και έτρεχα μια-μια σημειώνοντας τα αποτελέσματα σ' ένα excel file. Πριν προχωρήσω στο επόμενο μοντέλο, με κατάλληλους υπολογισμούς με φόρμουλες στην excel έβγαιναν και τα στατιστικά στοιχεία για το πόσα normal και πόσα abnormal έβαζε σωστά το δίκτυο, στην κατάλληλη κατηγορία. Οι κατηγορίες ήταν οι εξής 4:

- True positive (TP): Αριθμός των περιπτώσεων που το σύστημα βρήκε θετικές και είναι στη πραγματικότητα θετικές.
- False positive (FP): Αριθμός των περιπτώσεων που το σύστημα βρήκε θετικές και είναι στη πραγματικότητα αρνητικές.
- True negative (TN): Αριθμός των περιπτώσεων που το σύστημα βρήκε αρνητικές και είναι στη πραγματικότητα αρνητικές.
- False negative (FN): Αριθμός των περιπτώσεων που το σύστημα βρήκε αρνητικές και είναι στη πραγματικότητα θετικές.

Η ταξινόμηση σε κατηγορίες φαίνεται πως γίνεται και πιο κάτω:

Actual Condition	Classifier Prediction	Classification
1	1	True Positive
0	1	False Positive
1	0	False Negative
0	0	True Negative

Πίνακας 3.1

Κεφάλαιο 4

Περιγραφή λύσεων και αποτελέσματα

4.1 Αρχιτεκτονικές και αποτελέσματα	24
4.4.1 Αρχιτεκτονική με ένα κρυφό επίπεδο	26
4.1.2 Αρχιτεκτονική με δύο κρυφά επίπεδα	27
4.1.3 Αρχιτεκτονική με τρία κρυφά επίπεδα	28
4.2 Καλύτερα αποτελέσματα	32

4.1 Αρχιτεκτονικές και αποτελέσματα

Έγιναν λοιπόν δοκιμές σε πάρα πολλές αρχιτεκτονικές, μέχρι να βρω την αρχιτεκτονική που θα μου έδινε τα καλύτερα αποτελέσματα. Κατά τη διάρκεια των δοκιμών είχα την ευκαιρία να αλλάζω hidden layers με ένα, δύο ή τρία, να αλλάζω την πολυπλοκότητά τους αλλάζοντας τον αριθμό των νευρώνων στα hidden layers, καθώς επίσης και τις εποχές που θα έτρεχε το σύστημά μου. Μπορούσα επίσης να αλλάζω momentum και learning rate καθώς και weights αλλά προτιμούσα να τα κρατώ σταθερά για να έχω κάποιο μέτρο σύγκρισης κάθε αρχιτεκτονικής με τις υπόλοιπες, δηλαδή να μην συγκρίνω τελείως ανόμοια πράγματα.

Ακόμη σε μερικές αρχιτεκτονικές προσπάθησα να αφαιρέσω κάποια δεδομένα από τα δεδομένα εισόδου, ούτως ώστε να δω αν όλες οι είσοδοι παίζουν κάποιο ρόλο στο αποτέλεσμα. Αυτό που έκανα ήταν να αφαιρέσω την ηλικία της μητέρας που είναι ένας πολύ σημαντικός παράγοντας, σε κάποιες από τις αρχιτεκτονικές, και τα αποτελέσματα ήταν ακόμη πιο χαμηλά από τα προηγούμενα που έτρεξα κανονικά και με τις εννέα εισόδους. Το συμπέρασμα που έφτασα είναι ότι και οι εννέα είσοδοι που χρησιμοποιώ παίζουν η καθεμία τον δικό τους σημαντικό ρόλο στο αποτέλεσμα.

Στις συγκεκριμένες αρχιτεκτονικές που έτρεξα, υπήρχαν κάποιες παραμέτρους τις οποίες αν ήθελες είχες την ευκαιρία να τις αλλάξεις για να μπορούμε να έχουμε μια μεγάλη ποικιλία από συνδυασμούς. Κάποιες παραμέτρους που εξηγούνται αναλυτικά πιο κάτω, άλλαζα και κάποιες όχι για να μπορώ να έχω ένα μέτρο σύγκρισης.

Hidden slabs: είναι ο αριθμός των κρυφών επιπέδων που χρησιμοποιούνται στην αρχιτεκτονική που θέλω να τρέξω. Τα κρυφά επίπεδα μπορεί να είναι 1, 2 ή 3.

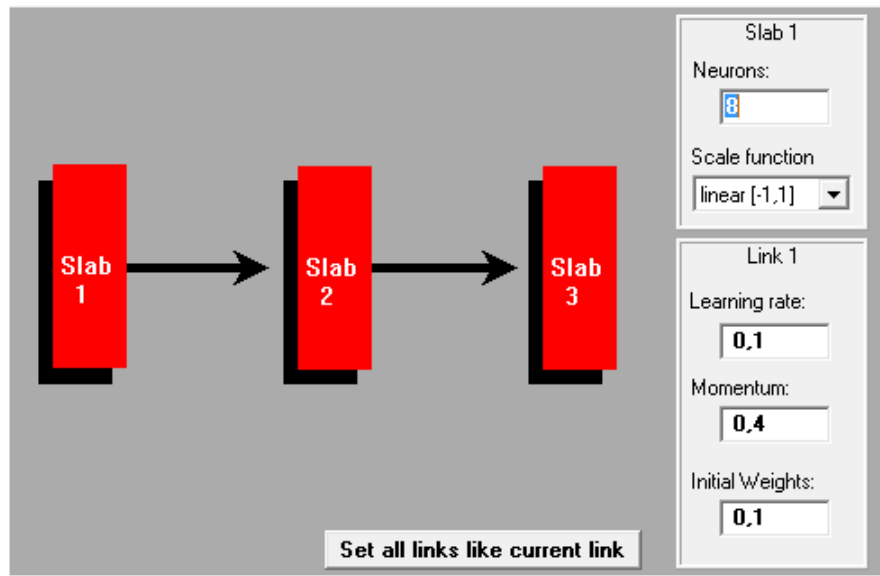
Momentum: κάποιες φορές η μάθηση δεν ολοκληρώνεται όπως πρέπει διότι δημιουργούνται κάποια προβλήματα (πρόβλημα των τοπικών ελαχίστων) και έτσι δεν καταλήγει το σύστημα μας στη καλύτερη δυνατή λύση. Με την ορμή (momentum) η μάθηση γίνεται γρηγορότερη, και ο παράγοντας αυτός λαμβάνει υπόψη το ποσοστό των αλλαγών των βαρών σε κάθε βήμα της εκπαίδευσης.

Learning Rate: χρησιμοποιείται για να αναπροσαρμοστούν τα βάρη. Χρησιμοποιείται στη συνάρτηση αναπροσαρμογής των βαρών μαζί με το σφάλμα που υπολογίζεται κάθε φορά.

Stopping criteria: ανάλογα με το πότε θέλουμε να σταματήσει το σύστημα να τρέχει, ρυθμίζουμε και το παράγοντα αυτό. Για παράδειγμα μπορούμε να χρησιμοποιήσουμε ένα συγκεκριμένο αριθμό εποχών.

Initial weights: αρχικοποιούμε τα βάρη τα οποία θα αλλάξουν στη συνέχεια ανάλογα με τη μάθηση του συστήματος.[3]

4.1.1 Αρχιτεκτονική με ένα κρυφό επίπεδο



Σχήμα 3.1

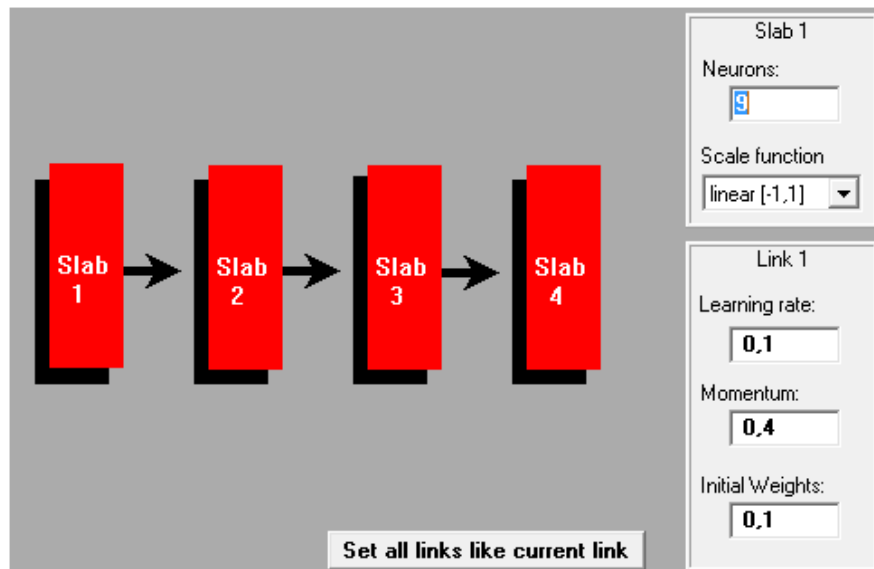
Σε αυτήν την αρχιτεκτονική που φαίνεται στο σχήμα 3.1 έτρεξα τα μοντέλα A1 μέχρι A12, τα μοντέλα A17, A18 καθώς και τα μοντέλα B10, B11, B12, B15 και B16.

Τα αποτελέσματα που έδωσε η αρχιτεκτονική αυτή ήταν από χαμηλά μέχρι πολύ μέτρια. Έφτασε μέχρι 57% στο training set και μέχρι 54% στο evaluation set. Εδώ δοκιμάστηκαν νευρώνες με πολυπλοκότητα από 20 μέχρι 120 και activation functions tanh και log. Μετά από σύγκριση, τα καλύτερα αποτελέσματα συμπεράναμε πως έδωσε η συνάρτηση της υπερβολικής εφαιτομένης tanh με σε συνδυασμό με πολυπλοκότητα νευρώνων 50 και 60, στο κρυφό επίπεδο.

Επιπλέον σε κάποιες από τις αρχιτεκτονικές, δοκίμασα να αφαιρέσω από τα δεδομένα εισόδου. Αφαίρεσα την ηλικία της κύησης κατά τη μέτρηση (Ga-d: Gestation age-days) για να δω πόσο ρόλο θα έπαιζε στην έξοδο του συστήματος το συγκεκριμένο δεδομένο. Παρατήρησα ότι το σύστημα παρουσίαζε υψηλότερα ποσοστά με εννέα εισόδους γι' αυτό και το επανέφερα στις εννέα αρχικές εισόδους (Βλέπε πίνακα 3.2 μοντέλα A10 με B10, A11 με B11 και A12 με B12)

4.1.2 Αρχιτεκτονική με δύο κρυφά επίπεδα

Τα μοντέλα που δοκιμάστηκαν σε αυτήν την αρχιτεκτονική είναι τα μοντέλα A13-A16, B15, B16 και επίσης τα μοντέλα C1-C13. Η αρχιτεκτονική φαίνεται στο σχήμα 3.2.

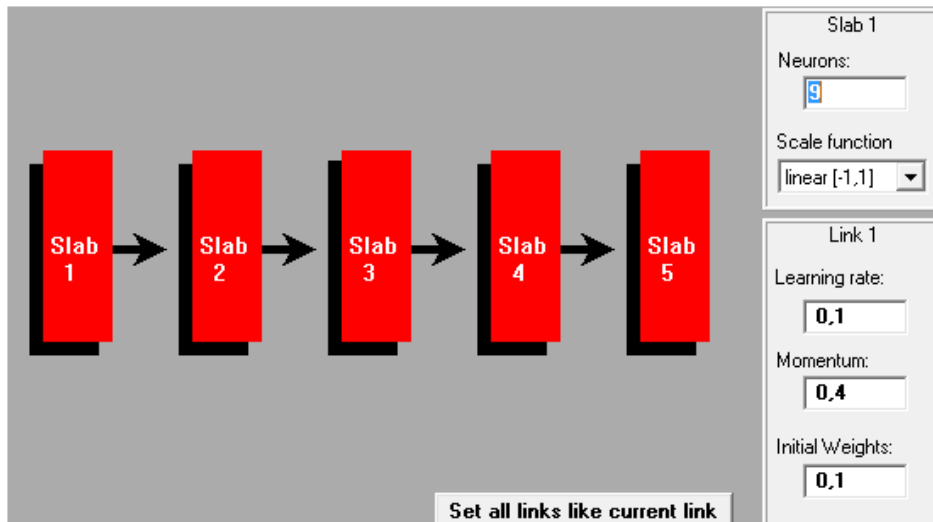


Σχήμα 3.2

Εδώ δοκιμάστηκε πολυπλοκότητα νευρώνων από 15 μέχρι 150 και activation functions tanh και log επίσης. Υπήρξε και μείωση των εισόδων κατά μια για να ελέγξω πάλι αν η ηλικία της κύησης έπαιξε κάποιο σημαντικό ρόλο, αλλά τα αποτελέσματα δεν ήταν και τόσο καλά και έτσι επαναφερθήκαμε στις εννέα αρχικές εισόδους και πάλι. Αυτό μπορεί να φανεί και από τον πίνακα 3.2 συγκρίνοντας τα μοντέλα A15 με το A16 και το B15 και B16 για παράδειγμα.

Τα υψηλότερα ποσοστά επιτυχίας που επιτεύχθηκαν στο δίκτυο αυτό ήταν 71% στο training set και 59% στο evaluation set. Τα ποσοστά αυτά έδωσε η αρχιτεκτονική που χρησιμοποιούσε επίσης την tanh σαν activation function και στα δύο layers και πολυπλοκότητα νευρώνων 60 στο πρώτο επίπεδο και 50 στο δεύτερο.

4.1.3 Αρχιτεκτονική με τρία κρυφά επίπεδα



Σχήμα 3.3

Στο δίκτυο αυτό δοκιμάστηκαν τα μοντέλα C14 μέχρι C20. Η πολυπλοκότητα των νευρώνων ήταν από 20 μέχρι 100 και η συνάρτηση ενεργοποίησης η tanh αφού όπως είδα και πιο πάνω μου έβγαζε τα καλύτερα αποτελέσματα.

Τα ποσοστά έφτασαν μέχρι το 75% στο training set και μέχρι το 72% evaluation set.

Πιο κάτω στον πίνακα 3.2 που χωρίζεται σε 3 μέρη φαίνονται τα αποτελέσματα από όλες τις αρχιτεκτονικές που έχω τρέξει, με όλες τις παραμέτρους που χρησιμοποιήθηκαν.

Πίνακας 3.2

NETWORK No	INPUTS	TOPOLOGY	LR	MO	WT	Epochs	TRAINING SET			EVALUATION SET		
							ALL %Corr	NORMAL %Corr	ABNORMAL L %Corr	ALL %Corr	NORMAL %Corr	ABNORMAL %Corr
A1	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-20log-1log		0.1	0.4	0.1	500	0.991112785	0.998929131	0.4902	0.991112785	0.998929131	0.49450549
A2	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-30log-1log		0.1	0.4	0.1	500	0.991489362	0.998814396	0.52206	0.991489362	0.998598726	0.53946154
A3	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-40log-1log		0.1	0.4	0.1	500	0.991602335	0.998890886	0.52451	0.993325778	0.998853503	0.51648352
A4	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-50log-1log		0.1	0.4	0.1	500	0.991564677	0.998814396	0.52696	0.993829493	0.999235669	0.52747253
A5	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-60log-1log		0.1	0.4	0.1	500	0.991037469	0.998814396	0.49265	0.992947991	0.998471338	0.51648352
A6	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-70log-1log		0.1	0.4	0.1	500	0.991638992	0.998929131	0.52451	0.993829493	0.999490446	0.50549451
A7	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-20tanh-1log		0.1	0.4	0.1	500	0.991903596	0.999005622	0.53676	0.993955421	0.999235669	0.53946154
A8	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-30tanh-1log		0.1	0.4	0.1	500	0.991978912	0.998967377	0.54412	0.994207279	0.999490446	0.53946154
A9	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-40tanh-1log		0.1	0.4	0.1	500	0.992129542	0.998929131	0.55637	0.993955421	0.999235669	0.53946154
A10	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-50tanh-1log		0.1	0.4	0.1	500	0.992506119	0.999082113	0.57108	0.993955421	0.99910828	0.54945055
A11	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-60tanh-1log		0.1	0.4	0.1	500	0.992506119	0.999043867	0.57353	0.993955421	0.99910828	0.54945055
A12	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-70tanh-1log		0.1	0.4	0.1	500	0.992506119	0.999120358	0.56863	0.99408135	0.999363057	0.53946154
A13	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-15log-15log-1log		0.1	0.4	0.1	500	0.992468462	0.998164225	0.62745	0.993325778	0.998343949	0.56043956
A14	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-25log-25log-1log		0.1	0.4	0.1	500	0.991790623	0.997705282	0.61275	0.99307392	0.997961783	0.57142857
A15	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-30log-30log-1log		0.1	0.4	0.1	500	0.992016569	0.998240716	0.59314	0.993199849	0.998089172	0.57142857
A16	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-50log-50log-1log		0.1	0.4	0.1	500	0.991414046	0.99747581	0.60294	0.99307392	0.997834395	0.58241758
A17	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-120log-1log		0.1	0.4	0.1	500	0.991150442	0.998737905	0.5049	0.992822063	0.998216561	0.52747253
A18	9:Gad.MA,Prev.CRL,NT.PAPP-A.b_hCf,FF: 9lin-60tanh-1log		0.1	0.4	0.1	500	a11	a11	a11	a11	a11	a11

B1	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-20log-1log	0,1	0,4	0,1	500				0,991790623	0,99877615	0,54412				0,993577635	0,998980892	0,52747253
B2	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-30log-1log	0,1	0,4	0,1	500				0,991338731	0,999005622	0,5				0,992947991	0,998598726	0,50549451
B3	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-40log-1log	0,1	0,4	0,1	500				0,991075127	0,998814396	0,4951				0,992947991	0,998598726	0,50549451
B4	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-50log-1log	0,1	0,4	0,1	500				0,991225758	0,998967377	0,4951				0,992947991	0,998598726	0,50549451
B5	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-60log-1log	0,1	0,4	0,1	500				0,991752966	0,999043867	0,52451				0,99408135	0,999617834	0,51648352
B6	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-70log-1log	0,1	0,4	0,1	500				0,991150442	0,998929131	0,49265				0,99307392	0,998853503	0,49450549
B7	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-20tanh-1log	0,1	0,4	0,1	500												
B8	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-30tanh-1log	0,1	0,4	0,1	500												
B9	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-40tanh-1log	0,1	0,4	0,1	500												
B10	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-50tanh-1log	0,1	0,4	0,1	500				0,992355489	0,999311585	0,54657				0,993955421	0,999235669	0,53846154
B11	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-60tanh-1log	0,1	0,4	0,1	500				0,992468462	0,999158603	0,56373				0,994207279	0,999235669	0,56043956
B12	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-70tanh-1log	0,1	0,4	0,1	500				0,9921672	0,998967377	0,55637				0,993325778	0,998726115	0,52747253
B13	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-15log-15log-1log	0,1	0,4	0,1	500												
B14	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-25log-25log-1log	0,1	0,4	0,1	500												
B15	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-30log-30log-1log	0,1	0,4	0,1	500				0,991941254	0,998049489	0,60049				0,992822063	0,997961783	0,54945055
B16	8:MA,Prev,CRL,NT,PAPP-A,b,hCG	FF: 8lin-50log-50log-1log	0,1	0,4	0,1	500				0,991225758	0,997781772	0,57108				0,992570205	0,997707006	0,54945055

C1	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-70log--70log-1log	0,1	0,4	0,1	0,1	500					0,991451704	0,999005622	0,50735			0,992570205	0,998598726	0,47252747
C2	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-80log--80log-1log	0,1	0,4	0,1	0,1	500					0,991150442	0,997667036	0,57353			0,992444277	0,997707006	0,53846154
C3	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-90log--90log-1log	0,1	0,4	0,1	0,1	500					0,990397289	0,998814386	0,45098			0,992318348	0,998980892	0,47158242
C4	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-100log--100log-1log	0,1	0,4	0,1	0,1	500					0,990349181	0,998393697	0,50735			0,992696134	0,998853503	0,46153846
C5	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-120log--120log-1log	0,1	0,4	0,1	0,1	500					0,991150442	0,99877615	0,50245			0,992947991	0,998980892	0,47252747
C6	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-150log--150log-1log	0,1	0,4	0,1	0,1	500					0,990058369	0,99812598	0,47304			0,991814633	0,997834395	0,47252747
C7	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-15log--15tanh-1log	0,1	0,4	0,1	0,1	500					0,991451704	0,999464566	0,47794			0,993703564	0,999617834	0,5
C8	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-30tanh--30tanh-1log	0,1	0,4	0,1	0,1	500					0,993560535	0,998623169	0,66912			0,993325778	0,998089172	0,5824758
C9	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-60tanh--60tanh-1log	0,1	0,4	0,1	0,1	500					0,993409904	0,99877615	0,64951			0,993955421	0,998343949	0,61538462
C10	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-80tanh--80tanh-1log	0,1	0,4	0,1	0,1	500					0,992995669	0,997896508	0,67892			1,0	0,99656051	0,62637363
C11	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-80tanh--40tanh-1log	0,1	0,4	0,1	0,1	500					0,993485219	0,998011244	0,70343			0,992947991	0,996815287	0,65934066
C12	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-60tanh--60tanh-1log	0,1	0,4	0,1	0,1	500					0,992769723	0,997169847	0,71078			0,991940562	0,99656051	0,59340859
C13	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-80tanh--40tanh-1log	0,1	0,4	0,1	0,1	1000					0,994840896	0,999464566	0,69853			0,993829493	0,998471338	0,59340859
C14	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-60tanh--60tanh-1log	0,1	0,4	0,1	0,1	500					0,9941254	0,998355452	0,72304			0,993451706	0,997579618	0,63736264
C15	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-60tanh--40tanh-1log	0,1	0,4	0,1	0,1	500					0,993221616	0,997055111	0,74755			0,992192419	0,996178344	0,64835165
C16	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-60tanh--60tanh-1log	0,1	0,4	0,1	0,1	500					0,995029185	0,99869966	0,7598			0,992696134	0,997324841	0,59340859
C17	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-70tanh--60tanh-1log	0,1	0,4	0,1	0,1	500					0,990397289	0,994071978	0,7549			0,987784914	0,990828025	0,72527473
C18	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-70tanh--60tanh-1log	0,1	0,4	0,1	0,1	500					0,9951045	0,99942632	0,71814			0,993325778	0,997579618	0,62637363
C19	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-30tanh--30tanh-1log	0,1	0,4	0,1	0,1	500					0,99363585	0,998049489	0,71078			0,992570205	0,997197452	0,59340859
C20	9;Gad;MA;Prev;CRL;NT;PAPP-A,b;hCf;FF;9lin-100tanh--100tanh-1log	0,1	0,4	0,1	0,1	500					0,993372246	0,998661414	0,65441			0,993829493	0,997834395	0,64835165

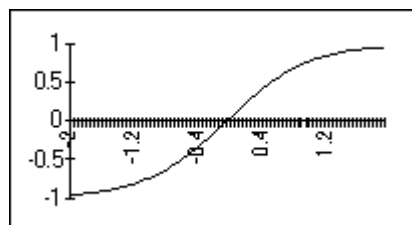
4.2 Καλύτερα αποτελέσματα

Τα καλύτερα αποτελέσματα φαίνονται στο abnormal % και normal % correlations όπου βλέπουμε τα ποσοστά επιτυχίας της κάθε αρχιτεκτονικής. Καλύτερα αποτελέσματα παρατηρούνται στις αρχιτεκτονικές C17, C15, C14 δηλαδή αρχιτεκτονικές με τρία κρυφά επίπεδα (hidden layers) και αρκετή πολυπλοκότητα σε σχέση με τον αριθμό των κρυφών νευρώνων. Τα καλύτερα αποτελέσματα φτάνουν μέχρι το 75,5% ποσοστό επιτυχίας στο training set και 72,5% στο evaluation set για την αρχιτεκτονική C17.

Η αρχιτεκτονική C17, είναι αρχιτεκτονική με τρία κρυφά επίπεδα, και με τα επίπεδα εισόδου και εξόδου έχουμε συνολικά πέντε επίπεδα. Ο αριθμός των νευρώνων με τα επίπεδα και τις συναρτήσεις ενεργοποίησης (activation functions) τους είναι οι εξής:

FF: 9lin-70tanh-60tanh-50tanh-1log

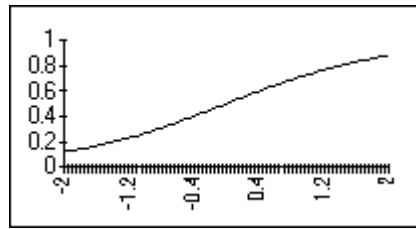
Η συνάρτηση ενεργοποίησης tanh[3], έχω να παρατηρήσω ότι όπου χρησιμοποιήθηκε, έχει δώσει πολύ καλά αποτελέσματα. Όπως αναφέρεται και στο help της Neuroshell2 είναι η «υπερβολική εφαπτομένη», η οποία σύμφωνα με τους ειδικούς δίνει πολύ καλά αποτελέσματα, αλλά δεν πρέπει να χρησιμοποιείται πάντα, αλλά ανάλογα με την περίπτωση. Η γραφική παράσταση της είναι η ακόλουθη:



Σχήμα 3.4

Tanh

Στην έξοδο έχουμε χρησιμοποιήσει την logistic activation function διότι οι εξοδοί μας είναι 0 ή 1 και η συνάρτηση αυτή προσδιορίζει πολύ καλά τις τιμές αυτές στο επίπεδο εξόδου. Πάντα χρησιμοποιείται η συνάρτηση αυτή όταν τα αποτελέσματα είναι σε κατηγορίες[3]. Η γραφική παράστασή της επίσης φαίνεται πιο κάτω:



Σχήμα 4.5

Logistic

Το momentum και learning rate, που χρησιμοποιήθηκαν στις αρχιτεκτονικές γενικά όπως και στη C17 είναι 0,4 και 0,1 αντίστοιχα. Τα βάρη αρχικοποιούνταν με 0,1 και οι εποχές που έτρεξε το σύστημά μας για να εκπαιδευτεί ήταν 500.

Μέσα στα δεδομένα που μου δόθηκαν, όπως φάνηκε και από τα αποτελέσματα, υπήρξαν κάποιες εγγραφές οι οποίες ήταν προβληματικές και κανένα από τα συστήματα μας δεν μπορούσε καθόλου να μάθει ή και να προβλέψει καθόλου τους συγκεκριμένους ασθενείς με αποτέλεσμα να έπεφτε απόλυτα έξω από την πρόβλεψη. Οι ασθενείς αυτοί είναι αυτοί που χαρακτηρίζονται από το σύστημα FP(false positive) ή FN(false negative) και παρουσιάζονται σχεδόν σε όλα τα συστήματα που τρέξαμε. Αυτό ίσως οφείλεται σε κάποιο λάθος που μπορεί να έγινε στην εξέταση του συγκεκριμένου ασθενή ή ακόμη και σε λάθος κατά τη διάρκεια της αντιγραφής των αποτελεσμάτων στον υπολογιστή του νοσοκομείου. Όλα τα αποτελέσματα από την ταξινόμηση που μου πρόσφερε το κάθε δίκτυο είναι αποθηκευμένα και μερικά παραδείγματα από τους προβληματικούς ασθενείς είναι οι εξής ταυτότητες:

110644

102185

110710

101102

108138

Σαφώς είναι πιο ασφαλές για την υγεία ενός εμβρύου, να βρεθεί από το σύστημα FP παρά να χαρακτηριστεί ως FN. Αν χαρακτηριστεί FP από το σύστημα σημαίνει ότι το βρήκε θετικό στην περίπτωση μιας γενετικής ανωμαλίας ενώ στην πραγματικότητα είναι αρνητικό. Αν χαρακτηριστεί από το σύστημα FN είναι ακόμα χειρότερα διότι σημαίνει ότι το σύστημα βρήκε το έμβρυο αρνητικό ενώ είναι θετικό. Τουλάχιστον στην πρώτη περίπτωση δίνεται μια αμφιβολία και παράλληλα η ευκαιρία να

σημειωθούν περεταίρω εξετάσεις για να βρεθεί το πρόβλημα που δεν υπάρχει στο τέλος, ενώ στην δεύτερη περίπτωση δίνεται ένας εφησυχασμός, ενώ στην πραγματικότητα το έμβρυο πάσχει από κάποια γενετική ανωμαλία.

Θα ήθελα επίσης να αναφέρω ότι η χρησιμοποίηση τριών κρυφών επιπέδων για την επίλυση του προβλήματός μας, είναι θεωρητικά λάθος. Σύμφωνα με το θεώρημα του Kolmogorov[11] όλα τα προβλήματα στα νευρωνικά δίκτυα μπορούν να λυθούν με μέχρι 3 ενεργά επίπεδα. Δύο κρυφά και ένα της εξόδου. Δεν είναι λάθος να χρησιμοποιήσουμε τρία κρυφά απλά είναι σαν να κάνουμε κάτι περισσότερο απ' αυτό που χρειάζεται. Η λύση δηλαδή θα μπορούσε να βρεθεί και με τα ίδια αποτελέσματα, και με δύο κρυφά επίπεδα, με ίσως αύξηση της πολυπλοκότητας των νευρώνων. Αυτό μπορεί να αποτελέσει θέμα σε μελλοντική μελέτη.

Κεφάλαιο 5

Συμπεράσματα και μελλοντική εργασία

5.1 Συμπεράσματα	35
5.2 Μελλοντική εργασία	36

5.1 Συμπεράσματα

Η πρόβλεψη μιας γενετικής ανωμαλίας σε ένα έμβρυο, από ένα ηλεκτρονικό σύστημα μπορεί να βοηθήσει πολύ ένα γιατρό στη διάγνωσή του αλλά και να βγάλει από την ταλαιπωρία και το κίνδυνο, την έγκυο γυναίκα και το μωρό της. Μέχρι σήμερα χρειάζεται να κάνουν κάποιες εξειδικευμένες εξετάσεις στο έμβρυο για να αποφανθούν αν το μωρό θα γεννηθεί φυσιολογικό ή όχι, και πολλές φορές μπορεί το έμβρυο να μπαίνει σε κίνδυνο, εφόσον γίνεται προγεννητικός έλεγχος, και μάλιστα στις πρώτες εβδομάδες της κύησης.

Όπως λοιπόν καταλαβαίνουμε η αποτυχία μιας πρόβλεψης, μπορεί να επιφέρει πολύ κακά αποτελέσματα, εφόσον κάποιες γυναίκες αδυνατούν να αναλάβουν την ευθύνη τέτοιων παιδιών, και οδηγούνται τελικά στη αποβολή τους. Άρα είναι πολύ σημαντική η πρόβλεψη, όταν παίζουμε με ανθρώπινες ζωές, και πρέπει να είναι ακριβής και τεκμηριωμένη από τον γιατρό. Γι' αυτόν τον λόγο λοιπόν, χρειαζόμαστε να έχουμε τα υψηλότερα ποσοστά επιτυχίας και το καλύτερο δυνατό σύστημα, πριν με την καλύτερη δυνατή έρευνα και τεκμηρίωση.

Κατά την εκπόνηση της διπλωματικής εργασίας αυτής δείξαμε ότι τελικά η τεχνητή νοημοσύνη και τα νευρωνικά δίκτυα, πιο συγκεκριμένα μπορούν να φανούν πολύ χρήσιμα ακόμη και στον τομέα της ιατρικής. Τα αποτελέσματα που εξάχθηκαν και τα ποσοστά επιτυχίας που επιτύχαμε, είναι πολύ ενθαρρυντικά για να συνεχίσουμε την έρευνα, να βγάλουμε ακόμη καλύτερα αποτελέσματα και ακόμη να επεκταθούμε και

στην επίλυση και άλλων σημαντικών προβλημάτων υγείας με την βοήθεια των νευρωνικών δικτύων. Οι «ηλεκτρονικοί γιατροί» σε κάποια χρόνια να μπορούν να βοηθούν τους πραγματικούς, και το έργο των τελευταίων να γίνεται ακόμη πιο εύκολο.

Η δική μου συνεισφορά στο πρόβλημα, ήταν μπορώ να πω κάποιας αξίας εφόσον δείχθηκε ότι με την χρησιμοποίηση τελικά των νευρωνικών δικτύων μπορεί να βρεθεί κάποια λύση στο πρόβλημα και να δημιουργηθεί ένα ιδανικό δίκτυο πρόβλεψης γενετικών ανωμαλιών σε έμβρυα με υψηλότερα ποσοστά επιτυχίας από τα αυτά που εξάγαμε από τη διπλωματική αυτή. Έτσι η τεχνολογία της πληροφορικής και ο συνδυασμός της χρήσης της από ένα γιατρό, μπορεί οπωσδήποτε να δημιουργήσει ένα πολύ μεγάλο βήμα στο τομέα της ανάπτυξης συστημάτων υγείας. Υπάρχουν πράγματα που σαφώς ένας γιατρός, που είναι πάνω από όλα άνθρωπος με αδυναμίες, δεν μπορεί να θυμάται. Εδώ έρχεται λοιπόν η επιστήμη μας, η επιστήμη της πληροφορικής για να βοηθήσει στις ανθρώπινες αδυναμίες, αφομοιώνοντας την εμπειρία και ενθουμούμενη η ίδια, όσα ο ανθρώπινος νους δεν μπορεί.

5.2 Μελλοντική εργασία

Η έρευνα που διεξαγάγαμε δεν μπορεί οπωσδήποτε να τελειώσει εδώ. Το θέμα είναι πολύ ενδιαφέρον και με πολλούς ορίζοντες ανάπτυξης. Στην Neuroshell2, το λογισμικό που με βοήθησε, υπάρχουν άπειροι συνδυασμοί με αρχιτεκτονικές, αριθμό επιπέδων που θέλεις να χρησιμοποιήσεις και στο καθένα μπορείς να τοποθετήσεις διαφορετικές συναρτήσεις ενεργοποίησης και πολυπλοκότητα νευρώνων. Μπορεί λοιπόν σε μελλοντική εργασία, κάποιος άλλος φοιτητής να καλύψει ένα άλλο φάσμα συνδυασμών, μέχρι να φτάσουμε στα καλύτερα δυνατά αποτελέσματα.

Κάτι άλλο που θα μπορούσε ακόμη να δοκιμαστεί είναι η χρησιμοποίηση περισσότερων από 8 ή 9 παραμέτρων που χρησιμοποίησα στις αρχιτεκτονικές μου. Υπάρχουν από τα διαθέσιμα δεδομένα προς είσοδο, 17 παράμετροι που μπορούν να χρησιμοποιηθούν. Ίσως να μπορούσε να γίνει και κάποιου είδους έρευνα σε συνεργασία με τους γιατρούς για να εισαγάγουμε στο σύστημα τις παραμέτρους με την μεγαλύτερη σημασία για την έξοδο του συστήματος η οποία είναι αυτή που μας ενδιαφέρει να είναι η πιο σωστή.

Επίσης σε μελλοντική εργασία μπορούν να χρησιμοποιηθούν οι ίδιες αρχιτεκτονικές σε παρόμοια προβλήματα και όχι μόνο στην πρόβλεψη γενετικών ανωμαλιών σε έμβρυα. Υπάρχουν στην ιατρική αλλά και σε πολλούς άλλους τομείς διάφορα προβλήματα πρόβλεψης, που μπορούν να επιλυθούν με τεχνητή νοημοσύνη και νευρωνικά δίκτυα.

Τέλος, θα μπορούσε να γίνει ένας συνδυασμός των νευρωνικών δικτύων με γενετικούς αλγόριθμους για να μπορέσουν μέσω αυτού του τρόπου να επιλεγθούν και να χρησιμοποιηθούν στο δίκτυό μας, μόνο οι βέλτιστες παράμετροι, που θα βοηθήσουν περισσότερο στην επίλυση του προβλήματός μας. Θα μπορούσε ακόμη, να γίνει άλλη μια προσπάθεια μείωσης των δεδομένων από φυσιολογικές εγκυμοσύνες γιατί το ποσοστό τους σε σύγκριση με τις παθολογικές είναι πολύ υψηλό με ως αποτέλεσμα να δυσκολεύεται το νευρωνικό δίκτυο να γενικεύσει.

Βιβλιογραφία

- [1] Kypros H. Nicolaides, “The 11–13+6 weeks scan,” Fetal Medicine Foundation, London 2004
- [2] Trisomy 18 Foundation
http://www.trisomy18.org/site/PageServer?pagename=whatist18_whatist
- [3] Ward Systems Group, Inc. Neuroshell2 – Εργαλείο Help
<http://www.wardsystems.com/manuals/neuroshell2/index.html?idxhowuse.htm>
- [4] C.N. Neocleous, K. Nikolaides, K. Neokleous, C.N. Schizas , “First Trimester Diagnosis of Trisomy-21 Using Artificial Neural Networks”, 2010.
- [5] C.N. Neocleous, K. Nikolaides, K. Neokleous, C.N. Schizas , “Artificial Neural Networks for non-invasive chromosomal abnormality screening of fetuses”, 2010.
- [6] Υλικό Διαλέξεων Συστήματα Υπολογιστικής Νοημοσύνης (ΕΠΛ 444), Τμήμα Πληροφορικής Τμήμα Πληροφορικής, Πανεπιστήμιο Κύπρου, Χρίστος Σχίζας
- [7] Υλικό Διαλέξεων Πληροφοριακών Συστημάτων Μάθησης (ΕΠΛ 442), Τμήμα Πληροφορικής, Πανεπιστήμιο Κύπρου, Χρίστος Χριστοδούλου
- [8] Pubmed Health-Down Syndrome
<http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001992/>
- [9] Genetic home reference-Turner Syndrome
<http://learn.genetics.utah.edu/content/disorders/whataregd/turner/>
- [10] WedMD- Triploid Syndrome
<http://children.webmd.com/triploid-syndrome>

- [11] Kolmogorov complexity
http://en.wikipedia.org/wiki/Kolmogorov_complexity