Contents

1 Abstract

Τα τυπωμένα ηλεκτρονικά είναι μια αναδυόμενη τεχνολογία που έχει τη δυνατότητα να καταστήσει δυνατή τη διείσδυση της πληροφορικής σε μια μεγάλη ποικιλία καταναλωτικών προϊόντων, χάρη στην υπό του cent κατασκευή τους και τους εύκαμπτους παράγοντες μορφής τους. Πολλές από τις προβλεπόμενες εφαρμογές που θα υποστηρίξουν έχουν να κάνουν με την ταξινόμηση των δεδομένων που συλλέγονται από τυπωμένους αισθητήρες για την εξαγωγή μιας χρήσιμης ιδιότητας σχετικά με το υπό μέτρηση αντικείμενο. Η εκτυπωμένη μηχανική μάθηση (ML) αναπτύσσεται προκειμένου να πραγματοποιούνται τέτοιου είδους ταξινομήσεις από δεδομένα αισθητήρων. Επειδή τα τυπωμένα ηλεκτρονικά έχουν πολύ υψηλότερες απαιτήσεις σε έκταση και κατανάλωση ενέργειας σε σύγκριση με τα παραδοσιακά ηλεκτρονικά, αυτά τα μοντέλα ML πρέπει να εκτελούνται σε περιβάλλον με πολύ περιορισμένους πόρους. Ευτυχώς, η ευκολία κατασκευής τυπωμένων κυκλωμάτων με τη χρήση προσθετικών μεθόδων επιτρέπει την πλήρη προσαρμογή του υλικού στο ακριβές εκπαιδευμένο μοντέλο που ενσωματώνει. Αυτό επιτρέπει τη συρρίκνωση των απαιτήσεων σε πόρους κατά πολλούς παράγοντες. Η παρούσα εργασία αξιολογεί τη δυνατότητα ανάπτυξης αρχιτεκτονικών δυαδικών νευρωνικών δικτύων (BNN) ως τυπωμένων ταξινομητών, όπου τα BNN είναι δίκτυα με βάρη και ενεργοποιήσεις που κβαντίζονται σε ένα μόνο bit για τη μείωση των υπολογιστικών απαιτήσεων στο ελάχιστο, γεγονός που τα καθιστά μια καλή υποψήφια αρχιτεκτονική για το συγκεκριμένο πρόβλημα.

2 Εισαγωγή

2.1 Εισαγωγή στην πανταχού παρούσα πληροφορική

Η τεχνολογία γενικά και πιο συγκεκριμένα ο υπολογισμός παίζει έναν ολοένα και μεγαλύτερο ρόλο στη ζωή μας και δεν υπάρχουν ενδείξεις ότι η τάση αυτή θα επιβραδυνθεί σύντομα. Εξακολουθεί ωστόσο να υπάρχει ακόμη ένα σχετικά άκαμπτο χάσμα μεταξύ του πραγματικού κόσμου και του υπολογιστικού τομέα, πράγμα που σημαίνει ότι οι περισσότερες από τις αλληλεπιδράσεις μας με τον κόσμο γύρω μας δεν εμπεριέχουν την πραγματοποίηση οποιουδήποτε υπολογισμού. Δεν είναι δύσκολο να φανταστούμε αμέτρητα παραδείγματα όπου υπολογιστικά στοιχεία θα προσέθεταν αξία σε καθημερινές δραστηριότητες όπως τα ψώνια από το παντοπωλείο ή θα μείωναν την απαιτούμενη εργασία σε διαδικασίες παραγωγής, όπως η μεταποίηση, εάν τα στοιχεία αυτά είχαν σχεδόν μηδενικό κόστος και μεγαλύτερη ενσωματωσιμότητα που συνδέεται με αυτά.

Αν και σχεδόν όλοι στις ανεπτυγμένες χώρες φέρουν μαζί τους και αλληλεπιδρούν με ισχυρούς υπολογιστές παντού όπου τυχαίνει να βρίσκονται, η μορφή της αλληλεπίδρασης δεν μπορεί εύκολα να προσαρμοστεί στο πλαίσιο όπου βρίσκονται. Δεν μπορεί κανείς απλά να ρωτήσει τις μπανάνες που πήρε αν είναι αρκετά ώριμες, να φωνάξει τα κλειδιά του για να βρει πού τα άφησε, να ελέγξει με τα παπούτσια του πόσα βήματα μπορούν να κάνουν ακόμα. Επιπλέον, είναι σαφές ότι οι μη μετρήσιμες διαδικασίες είναι τρομερά μη βελτιστοποιημένες σε σύγκριση με το τι θα μπορούσε να επιτευχθεί αν μια συνεχής ροή λεπτομερούς πληροφόρησης από κάθε ένα από τα στοιχεία που την απαρτίζουν και πρόσβαση σε έλεγχο της κάθε λεπτομέρειας αυτών των στοιχείων ήταν διαθέσιμη. Σκεφτείτε για παράδειγμα ένα αγρόκτημα όπου κάθε μεμονωμένος καρπός σε κάθε δέντρο έχει να παρακολουθείται η πρόοδος της ανάπτυξής του.

Παίρνοντας ουσιαστικά τις ιδέες του Διαδικτύου των Πραγμάτων (IoT) και προωθώντας στο λογικό τους όριο, η πανταχού παρούσα πληροφορική(ubiquitous computing) είναι ένα φιλόδοξο ιδανικό για ένα μέλλον όπου κάθε προϊόν είναι μια έξυπνη συσκευή, κάθε παρατηρήσιμη ιδιότητα, για την οποία οποιοσδήποτε θα μπορούσε λογικά να ενδιαφερθεί, είναι προσβάσιμη. Αυτοκίνητα θα είναι σε θέση να πλοηγούνται με ασφάλεια χωρίς πρόσβαση στην όραση ρωτώντας τις θέσεις των κοντινών συσκευών, αφού οτιδήποτε δεν είναι συσκευή άμεσα, έχει τουλάχιστον μία ή περισσότερες προσαρτημένες.

Τα τυπωμένα ηλεκτρονικά είναι σε θέση να διαδραματίσουν σημαντικό ρόλο τουλάχιστον στα πρώτα στάδια ενός τέτοιου μετασχηματισμού. Η εκτύπωση είναι επί του παρόντος η μόνη μέθοδος κατασκευής που μπορεί να παρέχει

υπολογιστικά στοιχεία κάτω του cent, και το κόστος είναι το μεγαλύτερο εμπόδιο για το πόσο διαδεδομένα μπορούν να γίνουν. Επιπλέον, η μη τοξικότητα είναι ζωτικής σημασίας για να καταστεί δυνατή η προσθήκη τους σε ταχέως κινούμενα καταναλωτικά αγαθά που είναι αναλώσιμα σε αυτές τις κλίμακες. Η ελαστικότητα βοηθά επίσης στην ευκολότερη ενσωμάτωση. Ακόμα και οι σχετικά "μέτριες" σε σύγκριση με το πλήρες όραμα εφαρμογές που μπορούμε να περιμένουμε ότι θα έρθουν τελικά, όπως οι ετικέτες RFID που αντικαθιστούν τους barcodes και επιτρέπουν στα καταστήματα να παρακολουθούν κάθε μεμονωμένο στοιχείο του αποθέματος ή οι τυπωμένοι αισθητήρες ποιότητας τροφίμων που καθιστούν τις ημερομηνίες λήξης παρωχημένες, έχουν μεγάλες δυνατότητες να διαταράξουν ένα ευρύ φάσμα βιομηχανιών.

Η μηχανική μάθηση μπορεί να επιταχύνει τη διαδικασία κατά πολλές τάξεις μεγέθους σε σύγκριση με το χρόνο που θα χρειάζονταν εξειδικευμένοι άνθρωποι σε πολλούς τομείς για να σχεδιάσουν ένα υπολογιστικό μοντέλο για να ερμηνεύσουν και να επεξεργαστούν τα δεδομένα των αισθητήρων. Σε πολλές περιπτώσεις, το αρχικό κόστος των μηχανικών θα ήταν αρκετό για να σταματήσει εντελώς η υιοθέτηση του προτύπου. Αν το μόνο που χρειάζεται είναι να συλλεχθούν και να επισημανθούν κάποια δεδομένα αισθητήρων, αυτό μπορεί εύκολα να αντιμετωπιστεί από οποιονδήποτε υπάλληλο. Εξαρτόμαστε επίσης από το ότι το autoML είναι αρκετά καλό για τις περισσότερες από αυτές τις εφαρμογές δεδομένων μικρής κλίμακας, καθώς διαφορετικά θα αντιμετωπίζαμε απλώς την ίδια δυσχέρεια με το να χρειαζόμαστε έναν επιστήμονα δεδομένων για κάθε μικρόπράγμα. Μέθοδοι για την ελάφρυνση των πόρων που απαιτούνται από το εκτυπωμένο σύστημα που υλοποιεί το μοντέλο, όπως η κβάντιση και η δυαδικοποίηση, μπορούν σαφώς να επεκτείνουν το εύρος του πόσο πολύπλοκη μπορεί να είναι η υποστηριζόμενη ταξινόμηση.

2.2 Εισαγωγή στα τυπωμένα ηλεκτρονικά

Τα τυπωμένα ηλεκτρονικά αναφέρονται σε πολύ λεπτές ηλεκτρονικές συσκευές και κυκλώματα που παράγονται με την εφαρμογή μελανιών με τις επιθυμητές ηλεκτρικές ιδιότητες σε διάφορα υποστρώματα. Μπορούν να κατασκευαστούν σε μεγάλο όγκο για πολύ χαμηλότερο κόστος σε σύγκριση με άλλα ηλεκτρονικά με μεθόδους που είναι κοινές στην βιομηχανία εκτύπωσης. Αυτό τα καθιστά ιδιαίτερα κατάλληλα για εφαρμογές όπου τα οφέλη της ηλεκτρονικής λειτουργικότητας από μόνα τους δεν αντισταθμίζουν τα σχετικά έξοδα. Επιπλέον, μπορούν να προσφέρουν ευέλικτες μορφές και τη δυνατότητα κάλυψης μεγάλης περιοχής. Ένα άλλο πλεονέκτημα που μπορεί να προκύψει από τη διάδοσή τους είναι η μείωση του αντίκτυπου των ηλεκτρονικών αποβλήτων, αφού τα τυπωμένα

ηλεκτρονικά μπορούν να είναι πολύ λιγότερο τοξικά για το περιβάλλον και πιο εύκολα ανακυκλώσιμα από τα υπόλοιπα, ή ακόμη και βιοδιασπώμενα. Δεν μπορούν να ανταγωνιστούν τα ηλεκτρονικά πυριτίου σε επιδόσεις λόγω της μεγάλης αντίστασης των αγώγιμων μελανιών, της έλλειψης υποστήριξης για υψηλές συχνότητες και της μεγάλης μεταβλητότητας κατά την κατασκευή. Ενώ η δυνατότητα να καλύπτουν μεγάλες περιοχές είναι μερικές φορές επιθυμητή, πολλές εφαρμογές απαιτούν σμίκρυνση που δεν μπορούν να προσφέρουν. Μια ποικιλία ενεργών και παθητικών συσκευών, όπως τρανζίστορ, αντιστάσεις, πυκνωτές, αισθητήρες, συλλέκτες και κεραίες μπορούν να υλοποιηθούν με αυτές. Θεωρούνται ότι αποτελούν μια αναδυόμενη αγορά με σημαντικές δυνατότητες διεύρυνσης του ρόλου των υπολογιστικών συστημάτων στην καθημερινή ζωή. Μπορούν να βοηθήσουν στη διείσδυση του Internet-of-Things να φτάσει πολύ βαθύτερα, και έτσι συνεργίζονται καλά με άλλες εξελίξεις στον τομέα. Μια πρόσφατη έκθεση της IDTechEx[1] προβλέπει την παγκόσμια αγορά για τα τυπωμένα εύκαμπτα ηλεκτρονικά, εξαιρουμένων των OLEDs, θα φθάσει τα 12 δισεκατομμύρια δολάρια έως το 2033.

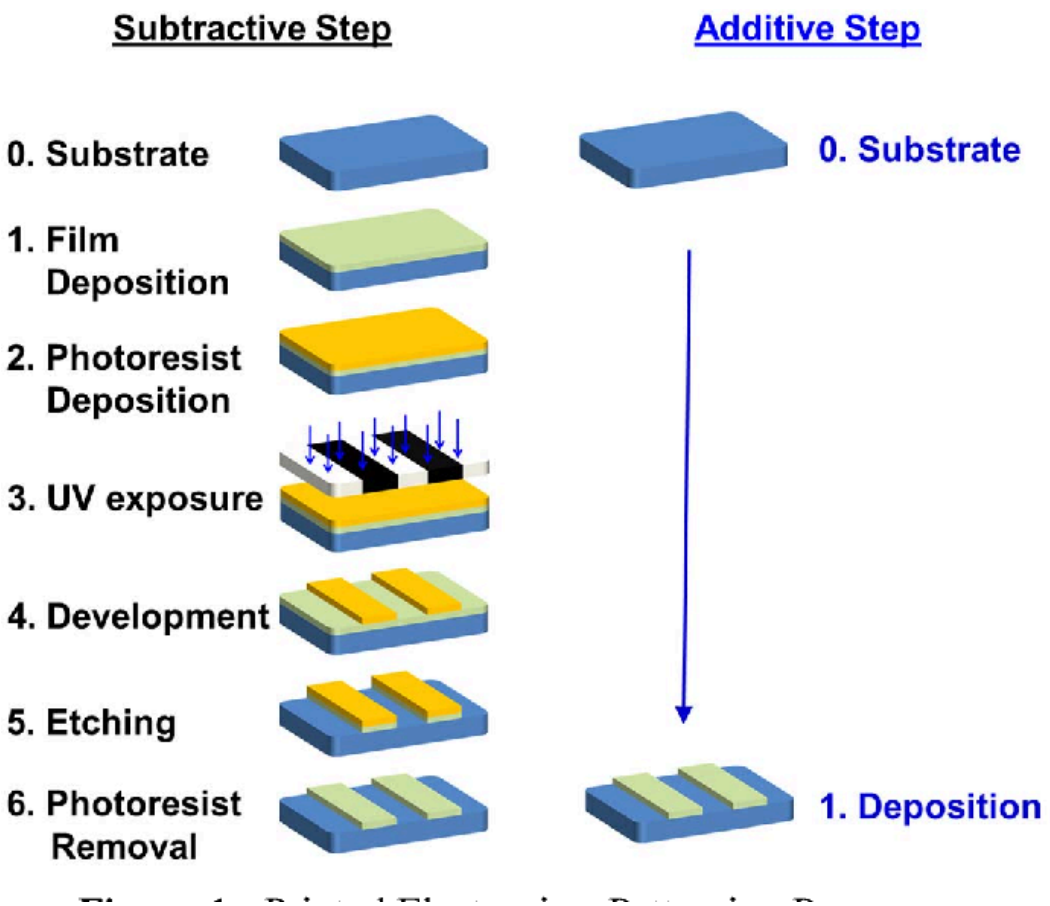

Figure 1: Comparison of subtractive electronics manufacturing to purely additive fabrication. The cost benefits of the much simpler additive procedure should be clear. Source: https://doi.org/10.1109/ISCAS.2017.8050614

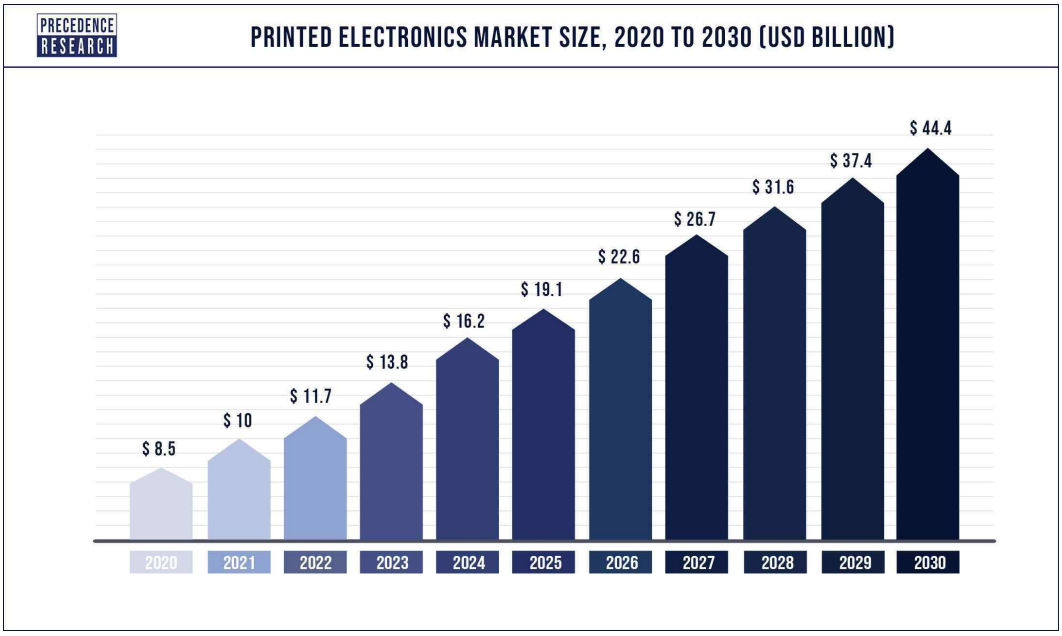

Figure 2: Projection of printed electronics market size. Source: Precedence Research

2.3 Χρήσεις των τυπωμένων ηλεκτρονικών

Η χρήση των τυπωμένων ηλεκτρονικών που οι περισσότεροι άνθρωποι μπορεί να γνωρίζουν στην καθημερινή τους ζωή είναι η μεμβράνη που χρησιμοποιείται για την ανίχνευση των πατημάτων των πλήκτρων στα περισσότερα μη μηχανικά πληκτρολόγια, ή ίσως οι αποπαγωτές παρμπρίζ.

Άλλες χρήσεις περιλαμβάνουν:

- Αισθητήρες: τα εύκαμπτα, βιοδιασπώμενα και ελαστικά αισθητήρια στοιχεία επιτρέπουν την αποτελεσματική παρακολούθηση πολλών διεργασιών. Μια ποικιλία ιδιοτήτων του κόσμου μπορεί να μετρηθεί με τυπωμένους αισθητήρες, όπως η θερμοκρασία, η αφή, η πίεση, τα αέρια, η υγρασία, τα επίπεδα φωτός και η παρουσία ορισμένων χημικών ουσιών. Η ευελιξία και η μη τοξικότητα είναι ιδιαίτερα σημαντικές για την ιατρική παρακολούθηση, γι' αυτό και οι βιοαισθητήρες έχουν λάβει μεγάλη προσοχή, ενώ ορισμένοι από αυτούς (π.χ. εκτυπωμένα επιθέματα ανίχνευσης επιληπτικών κρίσεων) είναι ήδη διαθέσιμοι στο εμπόριο.
- RFID: Το RFID (Radio Frequency Identification) είναι μια ασύρματη

τεχνολογία ανάγνωσης, η οποία επιτρέπει την απρόσκοπτη ταυτοποίηση και παρακολούθηση αντικειμένων μέσω μοναδικών κωδικών αναγνώρισης που αποθηκεύονται στις ετικέτες. Στόχος των τυπωμένων RFID είναι η αντικατάσταση των σημερινών μεθόδων αναγνώρισης αγαθών με έξυπνες ετικέτες. Οι ετικέτες RFID είναι συνήθως παθητικές και δεν απαιτούν παροχή ρεύματος. Μπορούν να κατασκευαστούν φτηνά με οποιαδήποτε κοινή μέθοδο εκτύπωσης. Έχει παρουσιαστεί ότι λειτουργούν σε συχνότητες 5G και WLAN και μπορούν να έχουν ακόμη και δυνατότητες αισθητήρων. Επί του παρόντος χρησιμοποιούνται κυρίως στα εισιτήρια και στην καταπολέμηση της κλοπής.

- Συγκομιδή ενέργειας: Οι τυπωμένες μπαταρίες μπορούν να παρέχουν ενέργεια στα λειτουργικά μέρη των τυπωμένων κυκλωμάτων μόνο για περιορισμένο χρονικό διάστημα και μπορεί να καταλαμβάνουν σημαντικό μέρος της επιφάνειας του κυκλώματος. Προκειμένου να καταστεί δυνατή η μεγαλύτερη αυτονομία στα αναπτυγμένα τυπωμένα ηλεκτρονικά συστήματα, η ικανότητα συγκομιδής ενέργειας από το περιβάλλον είναι ζωτικής σημασίας. Οι τυπωμένοι συλλέκτες μπορούν να αντλήσουν ενέργεια από ραδιοσήματα, δονήσεις και συνηθέστερα από το φως. Οι τυπωμένες φωτοηλεκτρικές/ηλιακές κυψέλες έχουν επίσης προσελκύσει μεγάλο ενδιαφέρον και εκτός του πεδίου των συγκομιστών για μικρά κυκλώματα, καθώς ενώ η απόδοσή τους δεν φτάνει τα επίπεδα των άκαμπτων ηλιακών κυψελών πυριτίου μπορούν να αναπτυχθούν σε μια ευρύτερη επιλογή χώρων, συμπεριλαμβανομένων των ενδυμάτων.

- Φωτισμός: Οι λυχνίες LED έχουν γίνει η κυρίαρχη πηγή φωτισμού, στη θέση των ενεργειακά σπάταλων λαμπτήρων πυρακτώσεως και των περιβαλλοντικών ναρκοπεδίων του φωτισμού με φθορίζοντες λαμπτήρες. Οι OLED αυξάνουν περαιτέρω την εξοικονόμηση ενέργειας και παράγουν πιο απαλό και ομοιόμορφο φωτισμό. Η εκτύπωση φαίνεται να αποτελεί μια πολλά υποσχόμενη λύση για την κατασκευή OLED χαμηλού κόστους με ανταγωνιστική φωτιστική απόδοση και να τους επιτρέπει να καλύπτουν μεγάλες επιφάνειες. Με αυτόν τον τρόπο έχουν καταδειχθεί φωτεινά πάνελ πάχους λεπτού χαρτιού.

- Οθόνες: Οι οθόνες είναι μία από τις πιο ώριμες πτυχές των τυπωμένων ηλεκτρονικών, με μεγάλες εκτυπωμένες οθόνες OLED 4K να είναι διαθέσιμες στο εμπόριο. Επιτρέπουν εύκαμπτες οθόνες, οι οποίες έχουν πολλές εφαρμογές στα ηλεκτρονικά είδη ευρείας κατανάλωσης και στα φορέσιμα και έτσι αποτελούν μια αγορά 5 δισεκατομμυρίων δολαρίων. Ακόμη και αν η εύκαμπτη οθόνη δεν είναι πλήρως τυπωμένη, τα τυπωμένα

ηλεκτρονικά μπορούν να της προσφέρουν πρόσθετα χαρακτηριστικά. Οι οθόνες QLED μπορεί επίσης μια μέρα να εκτυπωθούν, αν η ακρίβεια εκτύπωσης συνεχίσει να αυξάνεται.

- Wearables: Έξυπνες ηλεκτρονικές συσκευές που φοριούνται είναι ήδη πολύ δημοφιλείς, όπως έξυπνα ρολόγια ή ακουστικά βαρηκοΐας ή δαχτυλίδια NFC. Τα τυπωμένα ηλεκτρονικά έχουν πολλά να προσφέρουν στον χώρο χάρη στην ευελιξία τους. Έχουν αναπτυχθεί αγώγιμα υλικά που μπορούν να εκτυπωθούν σε ύφασμα και να αντέξουν στο πλύσιμο με απορρυπαντικό, επιτρέποντας την ενσωμάτωση ηλεκτρονικών σε κανονικά ρούχα. Οι εκτυπωμένοι αισθητήρες μπορούν να χρησιμοποιηθούν για την παρακολούθηση της δραστηριότητας, ένα από τα πιο δημοφιλή χαρακτηριστικά των σημερινών έξυπνων ρολογιών, ή για την παρακολούθηση της υγείας, με εκτυπωμένα επιθέματα για την ανίχνευση επιληπτικών κρίσεων που κυκλοφορούν ήδη στην αγορά. Μπορεί επίσης να φανταστεί κανείς ότι θα ενδιέφεραν τη βιομηχανία της μόδας.

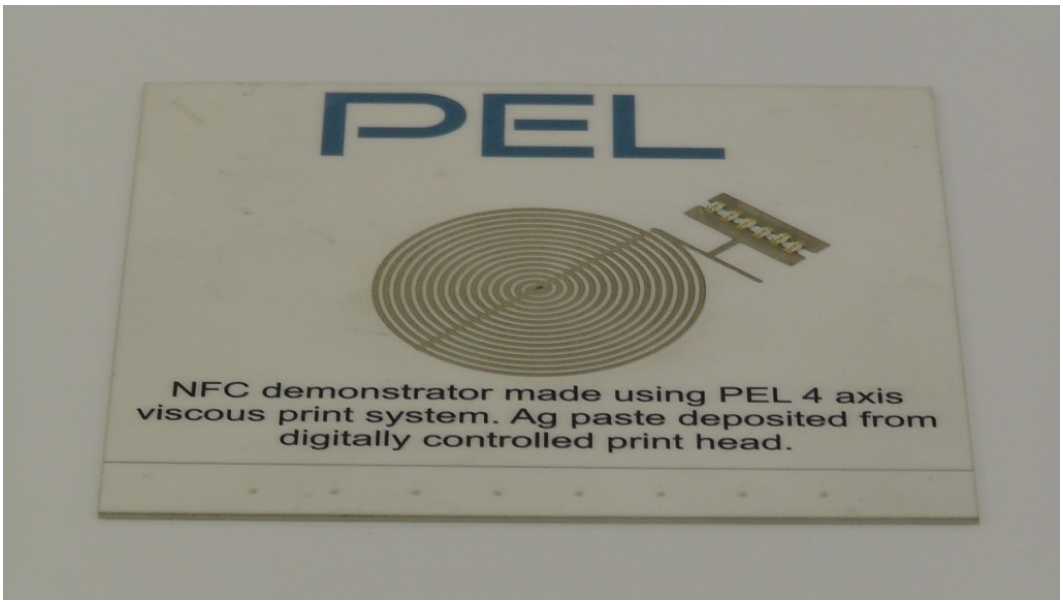

Figure 3: Printed NFC demontration. Source: PRINTED ELECTRONICS LTD

Figure 4: Printed circuit on the membrane of a common keyboard. Source: Paulo Maluf

2.4 TinyML

Το Edge Computing επιτρέπει εφαρμογές στις οποίες η επεξεργασία δεδομένων είναι ευαίσθητη στην τοποθεσία. Παρέχει μεγαλύτερες εγγυήσεις ασφάλειας, ιδιωτικότητας και διαθεσιμότητας στους τελικούς χρήστες. Αποτελεί θεμελιώδες στοιχείο της αγοράς IoT, που μπορεί να μειώσει την εξάρτηση από τα συστήματα cloud. Το κύριο εμπόδιο στην εξάπλωση της προσθήκης είναι οι περιορισμοί πόρων που επιβάλλει.

Για να αντιμετωπιστεί η απαίτηση εκτέλεσης εφαρμογών μηχανικής μάθησης στην ακμή για έξυπνες συσκευές, οι παραδοσιακές αρχιτεκτονικές είναι πολύ διογκωμένες για να ανταποκριθούν. Πολλά μοντέλα απαιτούν σήμερα υπολογιστικές δυνατότητες που δεν είναι εφικτές ακόμη και για το πιο high-end καταναλωτικό υλικό, πόσο μάλλον για συσκευές χαμηλής ισχύος. Το TinyML είναι ο τομέας της βελτιστοποίησης των αρχιτεκτονικών μηχανικής μάθησης ώστε να εκτελούνται σε συστήματα με εξαιρετικά περιορισμένους πόρους, συνήθως όχι περισσότερους από μερικά χιλιοστά του βατ.

Για το εγχείρημα αυτό απαιτείται διεπιστημονική εργασία, καθώς τόσο οι αλγόριθμοι ML, όσο και το λογισμικό και το υλικό που τους υποστηρίζει πρέπει να προσαρμόζονται σε αυτούς τους περιορισμούς, χωρίς να διακυβεύεται σε σημαντικό βαθμό η ακρίβεια των μοντέλων. Περίπου οι περιορισμοί που παίζουν ρόλο είναι η ενεργειακή απόδοση, η ικανότητα επεξεργασίας, ο χώρος μνήμης και το κόστος παραγωγής και σχεδιασμού. Θα πρέπει να τονιστεί ότι η ανησυχία αφορά το στάδιο inference της ML, αν και η δυνατότητα υλοποίησης της φάσης εκπαίδευσης σε υλικό άκρων αποτελεί επίσης μια δική της εξειδικευμένη προσπάθεια.

Δεδομένου ότι τα προ-εκπαιδευμένα μοντέλα ML δεν μπορούν να εκτελούνται σε αυτούς τους όρους κατά κανόνα, απαιτούνται pipelines από άκρη σε άκρη, από την απόκτηση δεδομένων έως την εξαγωγή συμπερασμάτων, δημιουργώντας τον τομέα της TinyML-as-a-Service ή TinyMaaS. Πρέπει να λαμβάνονται ειδικές προφυλάξεις σε κάθε ενδιάμεσο βήμα ώστε να οδηγηθούμε σε ένα αρκετά ελαφρύ μοντέλο εκτέλεσης.

Ορισμένες προσεγγίσεις του προβλήματος περιλαμβάνουν:

- Μια από τις πιο διαδεδομένες μεθόδους στο πεδίο είναι η αναζήτηση περιορισμένης νευρωνικής αρχιτεκτονικής (constrained neural architecture search, NAS). Η αναζήτηση νευρωνικής αρχιτεκτονικής εξετάζει ένα χώρο αναζήτησης διαφορετικών αρχιτεκτονικών για διαφορετικές υπερπαραμέτρους. Ένας αλγόριθμος προσπαθεί να εντοπίσει την καλύτερη

δυνατή αρχιτεκτονική για τη μεγιστοποίηση της απόδοσης του μοντέλου στην επίτευγματική συνάρτηση. Ένας αξιολογητής εξετάζει τους συμβιβασμούς μεταξύ ακρίβειας και αποδοτικότητας κατά την εγκατάσταση, δεδομένων των δηλωμένων περιορισμών μνήμης, ενέργειας κ.λπ. Μπορεί να εξετάσει ένα ή πολλά μοντέλα-στόχους σε μία ή περισσότερες πλατφόρμες. Τόσο ο χώρος αναζήτησης όσο και ο αλγόριθμος αναζήτησης έχουν επίγνωση του υλικού. Πρόκειται για ένα πρόβλημα βελτιστοποίησης πολλαπλών στόχων που συνήθως υλοποιείται ως πρόβλημα βελτιστοποίησης πολλαπλών σταδίων ενός στόχου. Η εκτέλεση της αναζήτησης είναι πολύ χρονοβόρα, αλλά τα αποτελέσματα υπερτερούν των περισσότερων χειροκίνητα σχεδιασμένων δικτύων.

- Μια προφανής προσέγγιση για τον χειρισμό του προβλήματος του περιορισμού μνήμης είναι η χρήση τεχνικών συμπίεσης δεδομένων στο μοντέλο ML. Μια βασική προσέγγιση που έχει επιδείξει συντελεστές συμπίεσης 15-40 φορές είναι τα προϊόντα Kronecker (KP). Ωστόσο, μπορεί να προκύψουν μεγάλες ποινές ακρίβειας και μια μέθοδος που ονομάζεται ντοπαρισμένο προϊόν Kronecker (DKP) αξιοποιεί την προσαρμογή των συν-μητρών για να προσπαθήσει να τις διορθώσει.

- Η υπερπαραμετροποίηση είναι η ιδιότητα ενός νευρωνικού δικτύου όπου οι πλεονάζοντες νευρώνες δεν βελτιώνουν την ακρίβεια των αποτελεσμάτων. Αυτός ο πλεονασμός μπορεί συχνά να αφαιρεθεί με μικρή ή καθόλου απώλεια ακρίβειας. Τα πλήρως συνδεδεμένα βαθιά νευρωνικά δίκτυα απαιτούν N^2 συνδέσεις μεταξύ των νευρώνων. Το κλάδεμα του δικτύου αφαιρεί παραμέτρους που δεν επηρεάζουν την ακρίβεια σε μεγάλο βαθμό. Μια συνηθισμένη περίπτωση όπου αυτό μπορεί εύκολα να γίνει είναι όταν οι παράμετροι είναι είτε μηδέν είτε αρκετά κοντά σε αυτό. Ομοίως όταν οι τιμές των παραμέτρων είναι περιττά επαναλαμβανόμενες. Μπορεί να εφαρμοστεί σε οποιαδήποτε κλίμακα, από μεμονωμένες συνδέσεις, νευρώνες έως ολόκληρα επίπεδα. Όταν μια διαδικασία κλαδέματος έχει ως αποτέλεσμα το νευρωνικό δίκτυο να χάσει τη συμμετρική του δομή αναφέρεται ως μη δομημένο κλάδεμα, διαφορετικά πρόκειται για δομημένο κλάδεμα. Το μη δομημένο κλάδεμα οδηγεί σε αραιούς πίνακες βαρών που οι γενικοί επεξεργαστές δεν εκτελούν αποτελεσματικά. Η επανεκπαίδευση ενός δικτύου μετά το κλάδεμα των παραμέτρων που δεν συνέβαλαν αρκετά μπορεί να του επιτρέψει να επιτύχει υψηλότερες ακρίβειες από ό,τι πριν. Ακόμα και το κλάδεμα ενός τυχαία αρχικοποιημένου δικτύου χωρίς εκπαίδευση πριν ή μετά μπορεί να οδηγήσει σε ευπρεπή ακρίβεια.

- Η απόσταξη γνώσης είναι μια διαδικασία εκπαίδευσης ενός μικρότερου,

ρηχότερου μαθητικού δικτύου για να αποδίδει κατά ένα μεγάλο μέρος τα logits εξόδου ενός μεγαλύτερου, ικανότερου δικτύου δασκάλου που έχει εκπαιδευτεί με ικανοποιητική ακρίβεια. Οι πιο προηγμένες παραλλαγές περιλαμβάνουν σύνολα μικρών δικτύων που το καθένα προσπαθεί να ταιριάξει με τα αποτελέσματα του συγκεντρωμένου συνόλου ή την αυτοαπόσταξη κατά την οποία τα ρηχότερα στρώματα ενός βαθύ νευρωνικού δικτύου προσπαθούν να μάθουν να ταιριάζουν με τα πιο περίπλοκα χαρακτηριστικά των βαθύτερων στρωμάτων. Εάν η διαφορά μεγέθους μεταξύ των δικτύων μαθητή και δασκάλου είναι πολύ μεγάλη, ένα ενδιάμεσου μεγέθους δίκτυο βοηθού δασκάλου λαμβάνει τις απαντήσεις του δασκάλου και στη συνέχεια τις αποστάζει στο μαθητή.

- Η κβάντιση είναι η διαδικασία μείωσης της αριθμητικής ακρίβειας των τιμών στο μοντέλο. Τα δίκτυα χρησιμοποιούν συνήθως αριθμούς κινητής υποδιαστολής 32 bit κατά τη διάρκεια της εκπαίδευσης[2]. Οι πιο συνηθισμένοι στόχοι κβαντισμού για αυτά είναι είτε ακέραιοι αριθμοί 8-bit είτε ακέραιοι αριθμοί 4-bit. Σε πολλές περιπτώσεις το δίκτυο δεν αξιοποιεί πλήρως αυτό το επίπεδο ακρίβειας. Η μείωση της ακρίβειας σε αυτές τις περιπτώσεις μπορεί να ανακουφίσει τον υπολογιστικό φόρτο που συνεπάγεται με αμελητέες θυσίες ακρίβειας. Η εκπαίδευση με επίγνωση της κβάντισης είναι μια διαδικασία κατά την οποία το δίκτυο πλήρους ακρίβειας εκπαιδεύεται εκ νέου στη μειωμένη μορφή. Διαδέχεται την κβάντιση του δικτύου μόνο μετά την ολοκλήρωση της διαδικασίας εκπαίδευσης. Όταν η μέθοδος φτάνει στα όριά της, η ακρίβεια μειώνεται σε 1 bit. Τα δίκτυα με ακρίβεια ενός μόνο bit ονομάζονται δυαδικά νευρωνικά δίκτυα(BNNs).

2.5 Εκτυπωμένη μηχανική μάθηση

Οι πιο ξεκάθαρα επιθυμητές εφαρμογές των υπολογισμών σε τυπωμένα ηλεκτρονικά συστήματα στον τομέα της πανταχού παρούσας πληροφορικής επικεντρώνονται γύρω από κάποια μορφή ταξινόμησης. Για κάθε τυπωμένο κύκλωμα που περιλαμβάνει έναν ή περισσότερους αισθητήρες απαιτείται ένα σύστημα για την ερμηνεία των εξόδων αυτού του αισθητήρα σε μια χρήσιμη μορφή, εκτός από τις τετριμμένες περιπτώσεις.

Η δημιουργία χειροποίητων υπολογιστικών μοντέλων για κάθε συνδυασμό αισθητήρων και περιπτώσεων χρήσης στις οποίες περιλαμβάνονται και η σχεδίαση αποδοτικών σχεδίων υλικού για την υποστήριξη του καθενός θα απαιτούσε ένα συντριπτικό ποσό εργατοωρών. Η μηχανική μάθηση παρέχει ένα κοινό μέσο τόσο για τη δημιουργία μοντέλων πρόβλεψης όσο και αρχιτεκτονικών για την υλοποίηση σε κυκλώματα ενός τεράστιου φάσματος αυτών των εφαρμογών. Υπάρχουν λοιπόν σαφή κίνητρα για την παροχή αυτών των μέσων.

Οι δυσκολίες σε αυτό το εγχείρημα πηγάζουν όλες από τους ακραίους περιορισμούς πόρων που δίνονται από τα μεγέθη των στοιχείων και τις ενεργειακές απαιτήσεις των τυπωμένων ηλεκτρονικών. Για να ξεπεραστούν αυτά τα εμπόδια, ολόκληρη η διαδικασία, από την επιλογή των κατάλληλων γενικών αρχιτεκτονικών μέχρι τη διατήρηση μόνο των απολύτως απαραίτητων στοιχείων υλικού στα κυκλώματα που προκύπτουν, πρέπει να βελτιστοποιηθεί για το σκοπό αυτό.

Πέρα από το να μειωθεί η επιφάνεια του κυκλώματος σε ένα λογικό μέγεθος, ένας σημαντικός παράγοντας που εμποδίζει τη χρηστικότητα ορισμένων υλοποιήσεων είναι ότι καμία τυπωμένη μπαταρία δεν μπορεί να υποστηρίξει την ηλεκτρική τους κατανάλωση. Επομένως, δεν είναι δυνατόν να τις έχουμε σε λειτουργία στην πράξη. Η ισχύς εξόδου αυτών των μπαταριών είναι ένα σκληρό όριο που πρέπει να επιτευχθεί ή η τυπωμένη ML είναι καθαρά θεωρητική. Στην παρούσα εργασία στοχεύεται η πιο παραχωρητική μπαταρία, της Molex, η οποία μπορεί να υποστηρίξει κυκλώματα με ισχύ έως και 30mW.

2.6 Στόχος διπλωματικής εργασίας

Η ιδέα αυτής της διατριβής λαμβάνει χώρα σε ένα σενάριο διάχυτου υπολογισμού. Έχω διασφαλίσει ότι η όλη διαδικασία από το σύνολο δεδομένων μέχρι τη netlist που μπορεί να περάσει στον εκτυπωτή δεν απαιτεί καμία χειροκίνητη παρέμβαση. Οποιοσδήποτε μπορεί να περάσει τα δεδομένα του αισθητήρα του στο ένα άκρο και να λάβει μετρήσεις για την ακρίβεια του μοντέλου, την επιφάνεια του κυκλώματος και τις απαιτήσεις ισχύος στο άλλο άκρο, χωρίς να απαιτούνται από αυτόν ειδικές γνώσεις σε οποιονδήποτε τομέα. Αυτό γίνεται συγκεκριμένα με τη χρήση εξειδικευμένων υλοποιήσεων δυαδικών νευρωνικών δικτύων, προκειμένου να αξιολογηθεί η αποτελεσματικότητά τους ως προς την παροχή μιας ραχοκοκαλιάς για τη διαδικασία αυτή.

Φανταστείτε, αν θέλετε, το σενάριο ενός ιδιοκτήτη καφετέριας. Αποφασίζει ότι θα ήθελε τα ποτήρια στα οποία σερβίρει τον καφέ του να υποδεικνύουν την ποσότητα ζάχαρης ή άλλων γλυκαντικών που χρησιμοποιούνται στο ρόφημα. Αυτό θα προλάβαινε τους πελάτες από το να αρπάζουν τον λάθος καφέ από το τραπέζι επειδή όλα φαίνονται δυσδιάκριτα. Αφού ψάχνουν σε ένα διαδικτυακό αποθετήριο για το ποιος αισθητήρας θα ήταν χρήσιμος εδώ, παραγγέλνουν μερικά δειγματοληπτικά φύλλα αυτών των εκτυπωμένων αισθητήρων και ένα μικρό gadget που κουμπώνει στο φύλλο και καταγράφει τις μετρήσεις των αισθητήρων. Αφού τους βυθίσουν σε δώδεκα καφέδες με διαφορετικά μίγματα γλυκαντικών μέσα, συνδέουν το gadget στον υπολογιστή τους και παίρνουν ένα spreadsheet με τις τιμές των αισθητήρων για κάθε περίοδο βύθισης. Απλώς προσθέτουν την ετικέτα που αποφάσισαν ότι αντιστοιχεί σε κάθε επίπεδο γλυκαντικού και περνούν το φύλλο στο σύστημα. Αποφασίζουν ότι η αναφερόμενη ακρίβεια και έκταση είναι διαχειρίσιμες και παραγγέλνουν το κύκλωμα που προκύπτει να τυπωθεί σε μια παρτίδα εύκαμπτων επιθεμάτων που μπορούν να επικολληθούν στο εσωτερικό των ποτηριών.

Έχουν αξιολογηθεί διάφορες αρχιτεκτονικές για ένα τέτοιο framework. Η παρούσα εργασία προσθέτει τα Δυαδικά Νευρωνικά Δίκτυα (BNN) στον κατάλογο των προσεγγίσεων για έντυπη εκτέλεση ML. Τα BNNs είναι δίκτυα που ποσοτικοποιούνται στο απόλυτο όριο του 1 bit. Έχουν σχεδιαστεί ειδικά για να ελαχιστοποιούν όσο το δυνατόν περισσότερο τους υπολογιστικούς πόρους και συνεπώς αποτελούν έναν βασικό υποψήφιο για να φέρουν όλο και περισσότερους ταξινομητές κάτω από το όριο υλοποίησης των τυπωμένων ηλεκτρονικών. Το μεγαλύτερο μέρος αυτής της διατριβής αφιερώνεται στην εξέταση αποδοτικών υλοποιήσεων υλικού προσαρμοσμένου κατά παραγγελία (bespoke) για τα BNNs ώστε να χωρέσουν στους περιορισμούς της τεχνολογίας.

3 Σχετικές εργασίες στη μηχανική μάθηση για τυπωμένα κυκλώματα

Από τότε που η τεχνολογία των τυπωμένων υπολογιστών έφτασε στο σημείο όπου τα μοντέλα Μηχανικής Μάθησης θα μπορούσαν να υποστηριχθούν για εκτέλεση, έχουν γίνει εργασίες για την υλοποίησή τους. Οι Tahoori et al [3] επιδεικνύουν έναν αναλογικό νευρώνα δύο εισόδων και δείχνουν πώς θα μπορούσε να επεκταθεί σε πλήρως τυπωμένα αναλογικά νευρωνικά δίκτυα με λειτουργίες MAC και ενεργοποίησης. Οι Douthwaite et al[4] χρησιμοποιούν κωδικοποίηση σημάτων στο πεδίο του χρόνου, αναπαριστώντας το μέγεθος ως πλάτος παλμού και κωδικοποιώντας τα βάρη με καθρέφτες ρεύματος. Η συσσώρευση γίνεται με γραμμική φόρτιση ενός πυκνωτή με τους κατοπτρισμένους παλμούς. Οι Gkoupidenis et al [5] μιμούνται βιολογικής έμπνευσης συναπτικές λειτουργίες με τρανζίστορ με ηλεκτρολυτική πύλη και δείχνουν πώς θα μπορούσαν να χρησιμοποιηθούν για ένα perceptron ενός στρώματος. Οι Ozer et al [6] οραματίζονται πώς θα μπορούσε να μοιάζει μια αυτόματη διαδικασία για τη δημιουργία εξειδικευμένων επεξεργαστών για μια ποικιλία αρχιτεκτονικών ML σε τυπωμένα ηλεκτρονικά, αλλά δεν προχωρούν πέρα από το στάδιο του οράματος. Οι Bleier et al [7] παρουσιάζουν έναν τυπωμένο μικροεπεξεργαστή με σύνολο εντολών προσαρμοσμένο στο εκάστοτε πρόγραμμα. Οι Weller et al [8] αξιοποιούν τον στοχαστικό υπολογισμό για να μειώσουν τις απαιτήσεις των μικτών αναλογικών-ψηφιακών νευρωνικών δικτύων, αλλά με βαρύ κόστος ακρίβειας.

Οι Mubarik et al [9] αξιολογούν μικρές αρχιτεκτονικές μηχανικής μάθησης (δέντρα απόφασης, τυχαία δάση και μηχανές διανυσμάτων υποστήριξης) σε ψηφιακές, βασισμένες σε πίνακες αναζήτησης και αναλογικές αρχιτεκτονικές σε ειδικά σχεδιασμένα τυπωμένα κυκλώματα. Εξετάζουν επίσης τις MLP, αλλά αποφασίζουν ότι είναι πολύ δαπανηρές για να τις αξιολογήσουν. Τα κυριότερα αποτελέσματα αφορούν τα δέντρα απόφασης (DT), όπου εξετάζουν τις απαιτήσεις των τυπωμένων υλοποιήσεων για βάθη από 1 έως 8, τόσο σε συμβατικά όσο και σε προσαρμοσμένα κυκλώματα. Δείχνουν ότι τα κατά παραγγελία(bespoke) κυκλώματα, τα οποία είναι μοναδικά κατάλληλα για τυπωμένα ηλεκτρονικά λόγω του χαμηλού μη επαναλαμβανόμενου μηχανικού(NRE) και κατασκευαστικού κόστους, μπορούν να υλοποιηθούν με περίπου δύο τάξεις μεγέθους χαμηλότερες απαιτήσεις από τα κυκλώματα που μπορούν να υποστηρίξουν ένα ευρύτερο φάσμα DTs και όχι μόνο ένα. Η παρούσα διατριβή είναι άμεσα εμπνευσμένη από αυτή την εργασία όσον αφορά τη χρήση σχεδιασμού κατά παραγγελία για τη μείωση των απαιτήσεων των υλοποιήσεων μοντέλων και εφαρμόζει κυρίως τις ιδέες τους στον τομέα των BNNs.

Οι Armeniakos et al [10] επεκτείνονται σε πιο απαιτητικά SVMs και Multi Layer Perceptrons. Προκειμένου να καταστεί δυνατή η υλοποίησή τους, αξιοποιούν τον προσεγγιστικό υπολογισμό με δύο τρόπους. Πρώτον, παρατηρούν ότι υπάρχει μεγάλη διακύμανση στις απαιτήσεις περιοχής ενός σταθερού πολλαπλασιαστή με βάση τον συντελεστή με τον οποίο πολλαπλασιάζει. Για παράδειγμα, ο πολλαπλασιασμός με μια δύναμη του δύο δεν απαιτεί καθόλου υλικό αφού πρόκειται για μια σταθερή μετατόπιση. Προσεγγίζουν τους συντελεστές βάρους των MLP και SVM για να εκμεταλλευτούν αυτή την παρατήρηση. Δεύτερον, εφαρμόζουν κλαδέματα μετά τη σύνθεση σε επίπεδο πύλης στη λίστα δικτύου των σχεδίων. Στοχεύουν σε πύλες που έχουν σχεδόν σταθερές εξόδους και επηρεάζουν μόνο τα λιγότερο σημαντικά bits των αποτελεσμάτων και τα αντικαθιστούν με τη σταθερή τιμή που ως επί το πλείστον εξάγουν. Μαζί αυτές οι προσεγγίσεις οδηγούν σε μειώσεις έκτασης και ισχύος περίπου κατά 2 φορές στις περισσότερες περιπτώσεις. Η εργασία αυτή είναι άμεση πηγή έμπνευσης για την παρούσα διατριβή, όπου οι συντελεστές βάρους ορίζονται στη φάση της εκπαίδευσης να είναι αποκλειστικά τιμές που δεν απαιτούν την υλοποίηση πολλαπλασιαστών, όπως συμβαίνει στα BNN. Έτσι, τα αποτελέσματα που επιτυγχάνονται εδώ συγκρίνονται με αυτά της εργασίας αυτής ως βάση. Η σύγκριση αυτή παρέχεται στην ενότητα Αποτελέσματα.

Στο επόμενο δημοσίευμα [11] εφαρμόζουν επιπλέον τεχνικές ελαχιστοποίησης των νευρωνικών συστημάτων, όπως κβάντιση, κλάδεμα και ομαδοποίηση βαρών, και τις συνδυάζουν χρησιμοποιώντας γενετικούς αλγορίθμους για να μειώσουν τις απαιτήσεις έκτασης έως και 8 φορές.

Στο [12], εκτός από τους προαναφερθέντες φιλικούς προς το υλικό συντελεστές και το κλάδεμα της λίστας δικτύων, εφαρμόζεται υπερκλιμάκωση τάσης(VOS) για την περαιτέρω μείωση των απαιτήσεων ισχύος των κυκλωμάτων ταξινομητών. Στη συνέχεια εφαρμόζεται ένας γενετικός αλγόριθμος για την ελαχιστοποίηση της περιοχής και τη μεγιστοποίηση της ακρίβειας για δεδομένο περιορισμό ισχύος. Αυτό επιτρέπει σε πολλά σχέδια να τροφοδοτούνται από τυπωμένες μπαταρίες θυσιάζοντας λιγότερο από 1% στην ακρίβεια.

Στη συνέχεια, το [13] επανεκπαιδεύει τα MLP με μια συνάρτηση βαθμολόγησης που λαμβάνει υπόψη το κόστος υλικού που σχετίζεται με τον πολλαπλασιασμό με τον συντελεστή κάθε βάρους. Οι συντελεστές ταξινομούνται σε συστάδες με βάση το κόστος υλικού τους και η επανεκπαίδευση επιτρέπει τη χρήση όλο και πιο ακριβών τιμών για τα βάρη μέχρι να επιτευχθεί το όριο ακρίβειας. Επιπλέον, τα προϊόντα αθροίζονται χρησιμοποιώντας προσεγγιστική πρόσθεση απορρίπτοντας τα λιγότερο σημαντικά bit των προϊόντων που συμβάλλουν λιγότερο στο αποτέλεσμα του MAC. Όλες αυτές οι βελτιώσεις οδηγούν σε 6πλάσια

εξοικονόμηση χώρου και ισχύος για απώλεια ακρίβειας 1% και 20πλάσια για 5%. Επειδή συχνά τα δίκτυά τους χρησιμοποιούν μόνο δυνάμεις του 2 ως βάρη και συνεπώς δεν χρησιμοποιείται υλικό για την εκτέλεση του πολλαπλασιασμού, αυτό το πλεονέκτημα της χρήσης BNN δεν υπάρχει εδώ. Ωστόσο, διαφορετικοί νευρώνες χρησιμοποιούν διαφορετικά βάρη για την ίδια είσοδο, με αποτέλεσμα λιγότερα ενδιάμεσα αθροίσματα να μπορούν να μοιραστούν μεταξύ των νευρώνων. Αυτό είναι ένα πλεονέκτημα που μπορούν να εκμεταλλευτούν τα BNNs, αν και πληρώνουν ένα τίμημα στις δυνατότητες αναπαράστασης.

Οι Balaskas et al στο [14] επεκτείνουν την ιδέα των φιλικών προς το υλικό συντελεστών στις τιμές κατωφλίου των συγκριτών στα δέντρα αποφάσεων. Πέρα από την τιμή κατωφλίου, η ακρίβεια της σύγκρισης μπορεί επίσης να διαμορφωθεί σε βάση ανά συγκριτή προκειμένου να αυξηθεί η αποδοτικότητα. Αναπτύσσουν έναν γενετικό αλγόριθμο για να βρουν βέλτιστες διαμορφώσεις των φιλικών προς το υλικό κατωφλίων κοντά στις αρχικές τιμές και μειωμένες ακρίβειες σύγκρισης χωρίς να θυσιάζεται ακρίβεια άνω του 1%. Ως αποτέλεσμα, η έκταση και η ισχύς μειώνονται κατά 3-4 φορές. Αυτό οδηγεί μερικά από τα μικρότερα σχέδια που εξετάζουν σε επιφάνεια κάτω από cm^2 και κατανάλωση ισχύος κάτω από mW.

Οι Iordanou et al [15] έχουν μια ενδιαφέρουσα προσέγγιση στην οποία χρησιμοποιούν γενετικό προγραμματισμό βασισμένο σε γράφους για να αναζητήσουν στο χώρο των λογικών εκφράσεων boolean αυτές που προβλέπουν την κλάση των δεδομένων του πίνακα με υψηλή ακρίβεια και μεταφράζοντας αυτές τις λογικές πύλες σε μια netlist. Το αποτέλεσμα είναι μια θάλασσα από λογικές πύλες, σε αντίθεση με τα δομημένα κυκλώματα άλλων προσεγγίσεων. Είναι περιττό να πούμε ότι αυτό απομακρύνεται από το πρότυπο των παραδοσιακών αρχιτεκτονικών ML στις οποίες τοποθετείται αυτή η εργασία.

4 Πληροφοριακό υπόβαθρο - Προαπαιτούμενα

4.1 Τεχνικές λεπτομέρειες για τα τυπωμένα ηλεκτρονικά

4.2 Μέθοδοι κατασκευής

Τα τυπωμένα ηλεκτρονικά κατασκευάζονται με τεχνικές από τη βιομηχανία γραφικών εκτυπώσεων. Διακρίνονται σε τεχνικές εκτύπωσης με επαφή ή R2R που χρησιμοποιούν πρότυπο και σε ανέπαφες που δεν χρησιμοποιούν. Απαιτούνται πολλαπλά βήματα εκτύπωσης για τα πολλαπλά στρώματα του κυκλώματος. Οι τεχνικές εκτύπωσης επαφής περιλαμβάνουν:

- Βαθυτυπία: Στη βαθυτυπία, ο κύλινδρος εκτύπωσης χαράσσεται με το πρότυπο και βυθίζεται εν μέρει στο μελάνι κατά τη διάρκεια της διαδικασίας, με μια λεπίδα να απορρίπτει το πλεονάζον μελάνι. Αυτό αφήνει μόνο μελάνι στα τμήματα του προτύπου, το οποίο μεταφέρεται στο υπόστρωμα υπό πίεση. Η βαθυτυπία μπορεί να εκτυπώσει με υψηλή ανάλυση και ταχύτητα σε σύγκριση με άλλες μεθόδους, αλλά το κόστος της χάραξης του κυλίνδρου την καθιστά χρήσιμη μόνο για μεγάλες παρτίδες.
- Offset: Στην εκτύπωση offset το σχήμα του προτύπου εναποτίθεται σε έναν κύλινδρο με μια ουσία που δέχεται μελάνι και το αρνητικό του προτύπου καλύπτεται με ουσίες που απωθούν το μελάνι. Με αυτόν τον τρόπο μόνο το σχήμα του προτύπου απορροφά μελάνι από έναν κύλινδρο μελάνης και στη συνέχεια μεταφέρεται στο υπόστρωμα μέσω ενός ενδιάμεσου κυλίνδρου.
- Φλεξογραφία: Το πρότυπο ενσωματώνεται σε μια εύκαμπτη πλάκα που τυλίγεται γύρω από έναν κύλινδρο εκτύπωσης έτσι ώστε τμήματα του σχήματος να υπερυψώνονται. Το μελάνι που εφαρμόζεται σε αυτόν τον κύλινδρο μεταφέρεται σε έναν δεύτερο κύλινδρο και στη συνέχεια στο υπόστρωμα, μόνο εάν βρίσκεται στα ανυψωμένα μέρη που αντιστοιχούν στο πρότυπο. Μπορεί να υποστηρίξει τόσο μη πορώδη όσο και πορώδη υποστρώματα.
- Μεταξοτυπία: Το "κόσκινο" σε αυτή την περίπτωση είναι ένα στενά συνδεδεμένο ύφασμα, έτσι ώστε το μελάνι να μπορεί να περάσει μόνο με την άσκηση πίεσης. Ένα στένσιλ του προτύπου τοποθετείται πάνω στο κόσκινο και μια λεπίδα ωθεί το μελάνι μέσω των ακάλυπτων τμημάτων στο υπόστρωμα. Η μεταξοτυπία είναι η απλούστερη τεχνική από όλες και μπορεί να δημιουργήσει παχύτερα στρώματα και να εκτυπώσει σε καμπύλες επιφάνειες. Πάσχει από χαμηλότερη ανάλυση σε σύγκριση με άλλες μεθόδους.

- Εκτύπωση με ταμπόν: Το μελάνι μπαίνει πάνω σε μια χάραξη του προτύπου. Στη συνέχεια, ένα μαλακό ταμπόν πιέζεται πάνω του και μεταφέρει το μελάνι με το επιθυμητό σχήμα στο υπόστρωμα. Μπορεί να εκτυπώσει σε επιφάνειες τρισδιάστατων αντικειμένων.

Οι τεχνικές χωρίς επαφή περιλαμβάνουν:

- Εκτύπωση μελάνης: Το μελάνι πέφτει πάνω στο υπόστρωμα από μικροσκοπικά στόμια. Είτε υπάρχουν αρκετά στόμια για να καλύψουν το πλάτος της εκτύπωσης είτε μπορούν να μετακινηθούν για να το κάνουν. Δεν απαιτεί μεγάλο εξοπλισμό και διαφορετικά σχέδια μπορούν να εκτυπωθούν σε υψηλή ανάλυση χωρίς επιπλοκές στην αλλαγή προτύπων, γεγονός που την καθιστά ιδανική για εκτύπωση κατά παραγγελία. Το κύριο μειονέκτημά της είναι η ταχύτητα εκτύπωσης. Η εκτύπωση συνεχούς ροής μελάνης έχει μια ροή μελάνης που κατευθύνεται πάνω στο υπόστρωμα ή σε έναν κάδο απόρριψης, ανάλογα με τις πληροφορίες του σχεδίου. Μπορεί να εκτυπώσει μεγαλύτερες παρτίδες από την Drop-on-Demand inkjet, αλλά με πέντε φορές χαμηλότερη ανάλυση. Η DoD ελέγχει αν θα ρέει μελάνι χρησιμοποιώντας μια βαλβίδα, ώστε να μην σπαταλιέται μελάνι. Αναπτύσσεται σε μικρότερες κλίμακες από τη συνεχή ροή.

- Αεροζόλ: Το μελάνι κονιορτοποιείται σε λεπτή ομίχλη μέσω πεπιεσμένου αέρα ή υπερήχων, επιταχύνεται και ψεκάζεται πάνω στο υπόστρωμα. Μπορεί να χρησιμοποιηθεί σε καμπύλες επιφάνειες και μπορεί να παρέχει ακόμη μικρότερα μεγέθη χαρακτηριστικών από ό,τι ο inkjet, αλλά είναι απαγορευτικά αργό.

Επιπλέον, μέθοδοι όπως η εναπόθεση υπό κενό, κατά την οποία εξατμιζόμενο μελάνι καλύπτει μια επιφάνεια σε κενό αέρος, ή η νανολιθογραφία με στυλό εμβάπτισης, κατά την οποία ένα μικροσκόπιο ατομικής δύναμης εφαρμόζει το μελάνι με μεγάλη ακρίβεια στο υπόστρωμα, θεωρούνται μερικές φορές ότι περιλαμβάνονται στην ομπρέλα των τυπωμένων ηλεκτρονικών, και παρόλο που μπορούν να επιτύχουν μικρότερα μεγέθη χαρακτηριστικών απαιτούν εξειδικευμένο εξοπλισμό και δεν είναι τόσο φιλικές προς το κόστος όσο οι παραδοσιακές μέθοδοι εκτύπωσης και συνεπώς λιγότερο συναφείς.

Figure 5: Dimatix DMP-2850 Materials Printer. Source: FUJIFILM

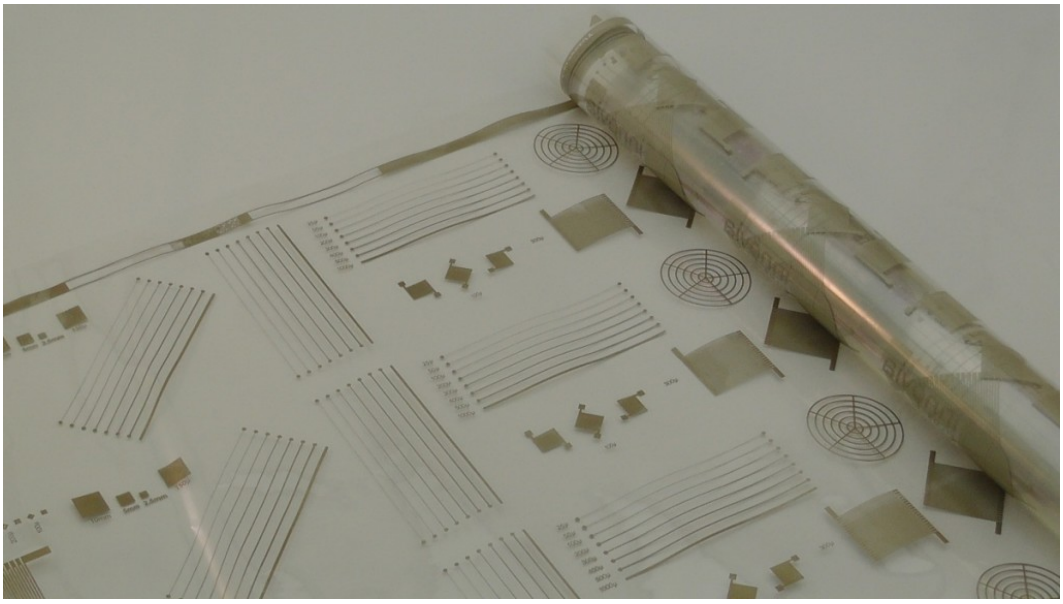

Figure 6: Long sheet of circuits printed on a Roll-to-Roll system. Source: PRINTED ELECTRONICS LTD

4.3 Μελάνια

Για την υλοποίηση λειτουργικών κυκλωμάτων απαιτούνται μελάνια με αγώγιμες, ημιαγώγιμες και διηλεκτρικές ιδιότητες. Συνήθως αποτελούνται από νανοσωματίδια υλικών με αυτές τις ιδιότητες που αναμειγνύονται με διαλύτη στο επιθυμητό ιξώδες και άλλα πρόσθετα για να καταστεί δυνατή η διαδικασία εκτύπωσης. Μπορούν να χρησιμοποιηθούν τόσο οργανικά όσο και ανόργανα υλικά.

- Αγώγιμα μελάνια: Η πλειονότητα των υλικών για αγώγιμα μελάνια είναι μεταλλικά νανοσωματίδια, με πιο συνηθισμένο το ασήμι. Αν και ο άργυρος ανήκει στην κατηγορία των πολύτιμων μετάλλων, το ασημένιο μελάνι δεν είναι φοβερά ακριβό, με στυλό με αγώγιμο ασημένιο μελάνι να κοστίζει λιγότερο από 4 δολάρια. Άλλα μέταλλα που χρησιμοποιούνται είναι ο χρυσός, το αλουμίνιο και ο χαλκός. Τα μελάνια χαλκού και αλουμινίου υφίστανται πολύ χειρότερη γήρανση από τα ασημένια. Τα οργανικά μελάνια βασίζονται συχνά σε νανοσωλήνες άνθρακα ή γραφένιο. Χρησιμοποιούνται επίσης φθηνότερα πολυμερή, παρά την κατώτερη αγωγιμότητά τους. Το πιο δημοφιλές είναι το PEDOT:PSS. Κεραμικά υλικά χρησιμοποιούνται επίσης σε αγώγιμα μελάνια, κυρίως οξείδιο του κασσίτερου του ινδίου(ITO), αν και πρόκειται για ακριβό υλικό.
- Ημιαγώγιμα μελάνια: Τα πιο συνηθισμένα ανόργανα υλικά που χρησιμοποιούνται είναι το πυρίτιο και το γερμάνιο και από τα οργανικά τα περισσότερα βασίζονται και πάλι σε CNT ή γραφένιο. Από αυτά μπορούν να παραχθούν τόσο υλικά τύπου p όσο και υλικά τύπου n, αν και οι τύποι p έχουν ιστορικά πολύ υψηλότερη απόδοση. (Το αντίθετο ισχύει για τα τρανζίστορ με πύλη ηλεκτρολύτη).
- Διηλεκτρικά μελάνια: Το διηλεκτρικό στρώμα πρέπει να είναι παχύτερο από το αγώγιμο και το ημιαγώγιμο στρώμα, ώστε να μην διαρρέει φορτίο μέσω αυτού. Ως ενεργό συστατικό μπορούν να χρησιμοποιηθούν υλικά υποστρώματος, κεραμικά οξείδια και πολυμερή.

4.3.1 EGFET

Η παρούσα εργασία βασίζεται στο Process Design Kit(PDK) για τρανζίστορ επίδρασης πεδίου οξειδίου με ηλεκτρολύτη(EGFET)[16]. Το EGFET χρησιμοποιεί στερεούς πολυμερείς ηλεκτρολύτες για την πύλη των τρανζίστορ αντί για διηλεκτρικά. Συγκεκριμένα χρησιμοποιείται οξείδιο του ινδίου για το σκοπό αυτό. Μπορούν να λειτουργήσουν σε συχνότητα έως και 250 Hz σε τάση 1V.

Το κύριο πλεονέκτημα των EGFET σε σύγκριση με τα οργανικά τρανζίστορ είναι ότι μπορούν να οδηγηθούν σε πολύ χαμηλές τάσεις, έως και 0,6V. Αυτό είναι ζωτικής σημασίας για την ικανοποίηση των περιορισμών που επιβάλλουν οι διαθέσιμες τυπωμένες μπαταρίες. Η περιοχή που καλύπτουν είναι ωστόσο σημαντικά μεγαλύτερη(10-100x) από εκείνη των οργανικών τρανζίστορ όπως τα CNT-TFT[17]. Για παράδειγμα, ένα απλό SR-latch που βασίζεται σε EGFET καταλαμβάνει $7mm^2$. Μια άλλη διαφορά των τρανζίστορ με ηλεκτρολύτη με τα αντίστοιχα οργανικά είναι ότι στο EGFET είναι δυνατόν να υλοποιηθούν μόνο τρανζίστορ τύπου n, ενώ στο CNT είναι δυνατόν να υλοποιηθούν μόνο τρανζίστορ τύπου p. Κανένα από τα δύο δεν υποστηρίζει και τα δύο, οπότε οι πύλες CMOS δεν μπορούν να χρησιμοποιηθούν σε τυπωμένα κυκλώματα.

4.4 Δυαδικά νευρωνικά δίκτυα

BNN είναι ο όρος για τα νευρωνικά δίκτυα που έχουν τόσο τις ενεργοποιήσεις όσο και τα βάρη με ακρίβεια 1 bit σε όλα τα κρυφά στρώματα. Τα στρώματα εισόδου οφείλουν να έχουν εισόδους υψηλότερης ακρίβειας, ώστε το δίκτυο να λαμβάνει επαρκείς πληροφορίες για να είναι δυνατή η ταξινόμηση, ενώ τα στρώματα εξόδου των ταξινομητών έχουν τις ενεργοποιήσεις τους να συγκρίνονται μεταξύ τους για να αποφασίσουν για την προβλεπόμενη κλάση, οπότε δεν μπορούν να δυαδικοποιηθούν. Ο πιο συνηθισμένος τομέας για τα BNN είναι τα Convolutional Neural Networks(CNNs). Παρουσιάστηκαν ανεξάρτητα το 2016 από τις εργασίες [18] και [19].

Πέρα από τη μείωση του μεγέθους αποθήκευσης που απαιτείται για τα βάρη $32\times$ σε σύγκριση με ένα δίκτυο πλήρους ακρίβειας 32-bit της ίδιας αρχιτεκτονικής, το υπολογιστικό κόστος μειώνεται επίσης σημαντικά, καθώς οι πράξεις πολλαπλασιασμού-συσσώρευσης(MAC) μπορούν να πραγματοποιηθούν με πράξεις XNOR και popcount. Αυτό μπορεί να οδηγήσει σε βελτίωση της ταχύτητας έως και $58\times$.

Κατά τη διάρκεια της εκπαίδευσης, χρησιμοποιούνται υποκείμενα βάρη υψηλότερης ακρίβειας για να γίνει η μάθηση πιο ισχυρή. Στη φάση προς τα εμπρός διάδοσης, αυτά τα ακριβέστερα βάρη, W, και οι ενεργοποιήσεις από το προηγούμενο επίπεδο I δυαδικοποιούνται χρησιμοποιώντας τη συνάρτηση προσήμου:

$$sign(x) = \begin{cases} +1, & \text{if } x \geq 0 \\ -1, & \text{if } x < 0 \end{cases}$$

Οι δυαδικές πράξεις $(-1, 1, *)$ και $(0, 1, \odot)$ είναι ισόμορφες, οπότε ο πολλαπλασιασμός των βαρών με τις ενεργοποιήσεις γίνεται με τη χρήση της πράξης XNOR όταν οι δυαδικές τιμές {-1,1} κωδικοποιούνται στις λογικές τιμές {0,1} για να αποθηκευτούν σε ένα bit.

Αυτή η απεικόνιση μπορεί να αναπαρασταθεί με τον γραμμικό μετασχηματισμό $f(x) = \frac{x+1}{2}$, αφού $f(-1) = 0$ και $f(1) = 1$. Η συσσώρευση και η επακόλουθη δυαδικοποίηση των προϊόντων ενεργοποίησης-βάρους $sign(I) * sign(W)$ μπορεί να υπολογιστεί με την εκτέλεση μιας πράξης popcount, η οποία επιστρέφει τον αριθμό των bit σε μια δεδομένη συλλογή που είναι 1, και τη σύγκριση του αποτελέσματος με ένα κατώφλι για τη δυαδικοποίηση.

Κατά τη διάρκεια της προς τα πίσω διάδοσης, πρέπει να χρησιμοποιηθεί μια προσέγγιση της συνάρτησης ενεργοποίησης $sign$, δεδομένου ότι το $sign$ έχει

παράγωγο 0. Η πιο συνηθισμένη μέθοδος είναι γνωστή ως straight-through estimation(STE). Στην STE η συνάρτηση πρόσημου προσεγγίζεται ως εξής:

$$STEsign(x) = \begin{cases} +1, & \text{if } x \geq 1 \\ x, & \text{if } 1 \geq x \geq -1 \\ -1, & \text{if } x \leq -1 \end{cases}$$

η οποία έχει παράγωγο:

$$\frac{dSTEsign}{dx} = \begin{cases} 1, & \text{if } 1 \geq x \geq -1 \\ 0, & elsewhere \end{cases}$$

Οι ενημερώσεις γίνονται στα υποκείμενα βάρη υψηλότερης ακρίβειας και οι δυαδικοποιήσεις τους χρησιμοποιούνται για το εμπρόσθιο πέρασμα.

4.5 Σύνολα δεδομένων

Τα σύνολα δεδομένων που επιλέχθηκαν για την εκπαίδευση των μοντέλων και την υλοποίηση είναι αυτά που χρησιμοποιήθηκαν από το [9]. Με αυτόν τον τρόπο τα αποτελέσματα για την ακρίβεια του μοντέλου και τις απαιτήσεις σε χώρο/ενέργεια μπορούν να συγκριθούν με άλλες προσεγγίσεις στη βιβλιογραφία. Όπως και σε αυτές τις εργασίες, τα κατηγορικά χαρακτηριστικά αφαιρέθηκαν από τα σύνολα δεδομένων, αφήνοντας μόνο τις εισόδους από αισθητήρες, αφού σε αυτές θα έχει πρόσβαση το πραγματικό εκτυπωμένο σύστημα (αυτή η υπόθεση μπορεί να παρακαμφθεί, αλλά αυτό είναι πέρα από το τρέχον αντικείμενο). Σημειώστε ότι η επιλογή των χαρακτηριστικών μπορεί να μην είναι η ίδια με τις προηγούμενες εργασίες, δεδομένου ότι τα κομμάτια δεδομένων που διατηρούσαν δεν τεκμηριώνονταν. Όλα τους ελήφθησαν από το αποθετήριο μηχανικής μάθησης του UCI[20].

Σύντομη περιγραφή των συνόλων δεδομένων:

- Arrythmia[21]: Διάγνωση καρδιακής αρρυθμίας από 12 ηλεκτροκαρδιογραφήματα.
- Cardio[22]: Διάγνωση προβλημάτων στον καρδιακό ρυθμό αγέννητων βρεφών.
- Pendigits[23]: Ταξινόμηση γραπτού ψηφίου από μια σειρά 8 σημάτων πίεσης από αισθητήρες αφής.
- Human Activity Recognition(HAR)[24]: Ταξινόμηση του τύπου της κίνησης ενός ατόμου (ορθοστασία, ανέβασμα σκαλοπατιών κ.λπ.) με τη χρήση επιταχυνσιόμετρων από κινητά τηλέφωνα στη μέση τους.

- Gas Id[25]: Ταξινόμηση της παρουσίας αερίου με χρήση χημικών αισθητήρων.
- Wine quality(white wine)[26]: Εκτίμηση της αντιλαμβανόμενης απόλαυσης των διαφόρων λευκών κρασιών με βάση την οξύτητα και τα ίχνη ανόργανων συστατικών.
- Wine quality(red wine)[26]: Αντίστοιχο με το παραπάνω για τα κόκκινα κρασιά.

Τα σύνολα δεδομένων χρησιμοποιούν εισόδους από αισθητήρες που αντιστοιχούν τουλάχιστον κατά προσέγγιση σε αισθητήρες που έχει αποδειχθεί ότι είναι δυνατόν να κατασκευαστούν με εκτύπωση. Το πλήρες σύστημα που περιλαμβάνει αισθητήρες, ταξινομητή και τροφοδοσία ρεύματος θα μπορούσε έτσι να υλοποιηθεί κάπως ρεαλιστικά και να μην απέχει πολύ από μια πραγματική περίπτωση χρήσης της τεχνολογίας.

Αισθητήρας	Dataset
Electrocardiography sensor on paper[27]	Arrythmia
Electrocardiography sensor on paper[27]	Cardio
Printed movement sensor	Human activity recognition
Printed gas sensor[28]	Gas identification
Printed piezoelectric sensor[29]	Pendigits
Printed pH sensor[30], Inkjet mineral sensor[31]	Wine Quality(White)
Printed pH sensor[30], Inkjet mineral sensor[31]	Wine Quality(Red)

5 Προτεινόμενο σύστημα

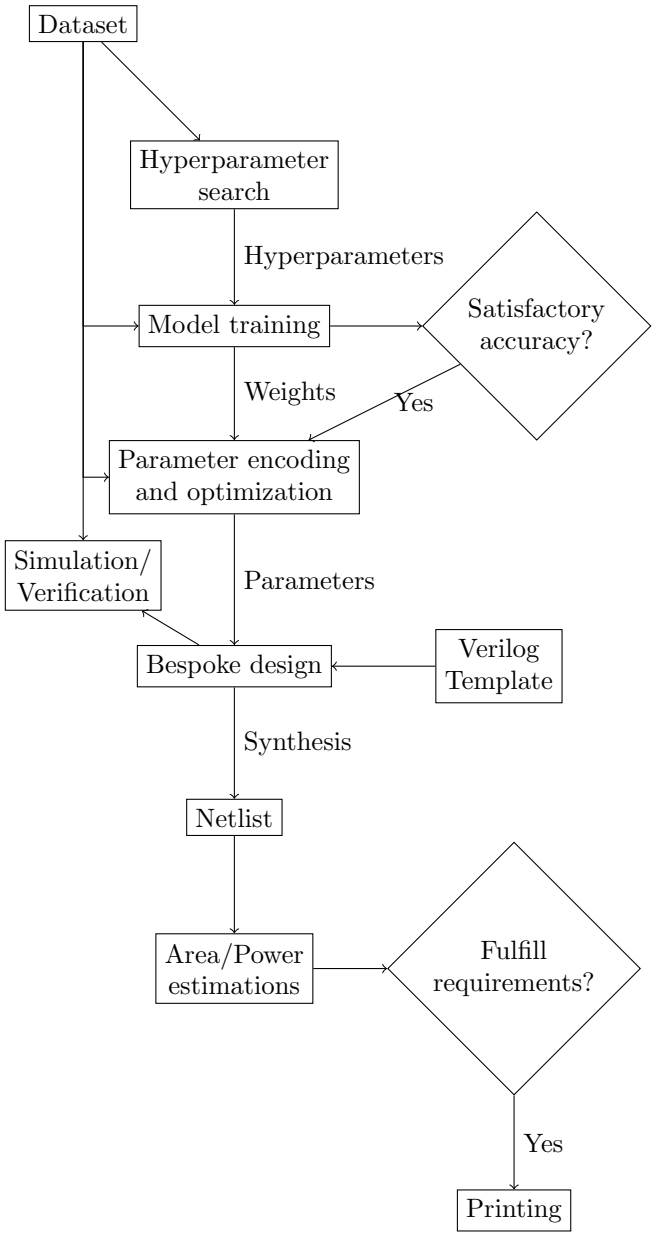

Figure 7: Προτεινόμενο σύστημα

Όπως περιγράφεται στην ενότητα πανταχού παρούσα πληροφορική, το παραδοτέο αυτής της διατριβής προορίζεται να είναι ένα πλαίσιο που επιτρέπει σε ένα

επισημασμένο σύνολο δεδομένων αισθητήρων να παράγει ένα πλήρως λειτουργικό τυπωμένο κύκλωμα που υλοποιεί έναν ταξινομητή για αυτό το σύνολο δεδομένων. Λόγω της έλλειψης πρόσβασης σε εξοπλισμό, η πραγματική εκτύπωση δεν είναι βιώσιμη σε αυτό το σημείο και το πεδίο εφαρμογής θα περιοριστεί στο μέρος της διαδικασίας από το σύνολο δεδομένων στη netlist.

Η αρχιτεκτονική του ταξινομητή θα είναι ειδικότερα ένα δυαδικό νευρωνικό δίκτυο (BNN), όπως εξηγείται στα προκαταρκτικά. Μπορεί να μην είναι απαραίτητα η καταλληλότερη αρχιτεκτονική σε κάθε περίπτωση, αλλά δημιουργήθηκε με σκοπό τη μείωση της κατανάλωσης πόρων, η οποία αποτελεί το βασικό μέλημα. Επιλέχθηκε έτσι ως η γωνιά που θα εξετάσω εδώ.

Η πρόχειρη ακολουθία των διεργασιών που εκτελούν τον εν λόγω μετασχηματισμό περιλαμβάνει τα εξής:

1. Αναζήτηση υπερπαραμέτρων: Προκειμένου μια διαδικασία εκπαίδευσης που δεν απαιτεί χειρισμό να είναι εφαρμόσιμη σε έναν αχανή χώρο πιθανών συνόλων δεδομένων αισθητήρων για τα οποία μπορεί να είναι επιθυμητοί οι εκτυπωμένοι ταξινομητές, μια ενιαία γενική διαμόρφωση των υπερπαραμέτρων εκπαίδευσης δεν επαρκεί. Έτσι, πρέπει να πραγματοποιηθεί αναζήτηση για την εύρεση ενός συνόλου κατάλληλου για την εκάστοτε κατανομή δεδομένων.
2. Εκπαίδευση μοντέλων: Αφού βρεθεί μια χρήσιμη διαμόρφωση υπερπαραμέτρων, ένα μοντέλο εκπαιδεύεται βάσει αυτών και αξιολογείται. Εάν η ακρίβεια του διαχωρισμού δοκιμής είναι επαρκής για τις ανάγκες του χρήστη, μπορεί να πραγματοποιηθεί η υπόλοιπη διαδικασία. Τα βάρη και η αρχιτεκτονική του δικτύου του τελικού μοντέλου διαβιβάζονται προς επεξεργασία.
3. Βελτιστοποίηση των παραμέτρων: Ορισμένες πληροφορίες που προκύπτουν από τα βάρη και το σύνολο δεδομένων μπορούν να χρησιμοποιηθούν για να βοηθήσουν τον σχεδιασμό να αποφύγει τους περιττούς υπολογισμούς. Για ένα ακραίο αλλά πρακτικά ισχύον παράδειγμα, εάν ένας νευρώνας δεν αλλάζει ποτέ την έξοδό του με βάση τις ενεργοποιήσεις εισόδου του, μπορεί να επισημανθεί ώστε να αντικατασταθεί με μια στατική σταθερή ανάθεση.
4. Κωδικοποίηση παραμέτρων: Τα βάρη, τα μέτρα της αρχιτεκτονικής του δικτύου και οι παράγωγες πληροφορίες υποβοήθησης είτε χρησιμοποιούνται για την παραγωγή κώδικα verilog που εκτελεί τους υπολογισμούς που συνεπάγονται είτε μορφοποιούνται έτσι ώστε να μπορούν να διαβαστούν και να αναλυθούν από τους μηχανισμούς ελέγχου της Verilog.
5. Ενσάρκωση της σχεδίασης: Ένα template για τον τύπο της σχεδίασης που θα έπρεπε να παραχθεί λαμβάνει τα αποσπάσματα της προσαρμοσμένης λειτουργικότητας ή/και τις μορφοποιημένες παραμέτρους που απαιτούνται

για την εξαγωγή της εν λόγω λειτουργικότητας με τη χρήση των generate blocks που εισάγονται σε αυτό. Το αποτέλεσμα είναι ένας σχεδιασμός που είναι προσαρμοσμένος στο ακριβές εκπαιδευμένο υπό εξέταση μοντέλο.

6. Επαλήθευση: Πραγματοποιείται προσομοίωση συμπεριφοράς του σχεδιασμού για να επιβεβαιωθεί ότι η ακρίβεια του ταξινομητή διατηρείται ικανοποιητικά.
7. Σύνθεση: Μια βελτιστοποιημένη λίστα δικτύου παράγεται από τις προδιαγραφές HDL. Πραγματοποιείται προσομοίωση σε επίπεδο πύλης για να εξασφαλιστεί η λειτουργικότητα.
8. Εκτίμηση μετρικών: Οι εκτιμήσεις εμβαδού και ισχύος λαμβάνονται από τα εργαλεία σύνθεσης και προσομοίωσης χρησιμοποιώντας πληροφορίες από το PDK των τυπωμένων εξαρτημάτων. Εάν οι απαιτήσεις αυτές φαίνεται να υποστηρίζονται από τον προϋπολογισμό της περίπτωσης χρήσης, ο χρήστης μπορεί να δώσει εντολή εκτύπωσης.

Τμήματα της ευρύτερης διαδικασίας που δεν εμπίπτουν στο πεδίο εφαρμογής του παρόντος πλαισίου είναι:

- Διαθεσιμότητα εκτυπωμένων αισθητήρων
- Πρόσβαση σε δεδομένα εκτυπωμένων αισθητήρων για την επισήμανση
- Εξέταση άλλων υποσχόμενων αρχιτεκτονικών
- Σχεδιασμός μασκών για την τοποθέτηση τυπωμένων στοιχείων
- Συμπερίληψη της κατανάλωσης πόρων αισθητήρα, ADC και ένδειξης εξόδου σε αναφερόμενες εκτιμήσεις.

Επιπλέον, το πλαίσιο παρέχει έναν τρόπο γρήγορης δοκιμής, αξιολόγησης και σύγκρισης υλοποιήσεων μοντέλων Verilog. Από τη στιγμή που έχει γραφτεί ο πρότυπος σχεδιασμός δεν απαιτούνται χειροκίνητα βήματα από τον σχεδιαστή για την εφαρμογή του σε κάθε μοντέλο που τον ενδιαφέρει, την επαλήθευση της λειτουργικότητάς του και τη λήψη feedback σχετικά με το πώς η νέα προσέγγιση συγκρίνεται με τις προηγούμενες. Αυτή η πτυχή του πλαισίου επέτρεψε στον πειραματισμό για την αναζήτηση αποδοτικών υλοποιήσεων BNN να προχωρήσει με ρυθμό που δεν θα ήταν εφικτός εάν τα βήματα αυτά δεν ήταν αυτοματοποιημένα.

Το υπόλοιπο της διατριβής θα ασχοληθεί σχεδόν εξ ολοκλήρου με τη σχεδίαση αποδοτικού υλικού ταξινομητή BNN κατά παραγγελία. Αυτό περιλαμβάνει τα μέρη 3 και 5 της διαδικασίας που αναφέρονται παραπάνω. Τα υπόλοιπα, αν και είναι χρονοβόρα για την υλοποίηση, δεν έχουν μέρη που να παρουσιάζουν ενδιαφέρον και θα συνοψιστούν στην ενότητα πειραματικής διάταξης.

Τα σημεία που νομίζω ότι είναι χρήσιμο να παρουσιαστούν εκ των προτέρων, προκειμένου να είναι κατανοητή η συζήτηση που θα ακολουθήσει για τις

υλοποιήσεις HDL, είναι τα εξής:

- Κατά τη διάρκεια της εργασίας στα σχέδια που παρουσιάζονται χρησιμοποιούνται σταθερά 6 μοντέλα που αντιστοιχούν σε 6 από τα 7 σύνολα δεδομένων που παρουσιάζονται παραπάνω για τη δοκιμή και τη σύγκριση των αποτελεσμάτων. Το ένα σύνολο δεδομένων από τα 7 που δεν κατάφερε να περάσει ήταν το Arrhythmia, επειδή η εκμαθημένη στρατηγική του μοντέλου δεν ήταν αποδεκτή.

- Όλα αυτά τα δίκτυα έχουν ένα κρυφό στρώμα που δέχεται εισόδους 4 bit που λαμβάνονται από τους ADC που είναι συνδεδεμένοι με τους αντίστοιχους αισθητήρες και ένα στρώμα εξόδου που δέχεται δυαδικές εισόδους 1 bit από το προηγούμενο στρώμα και παράγει μια βαθμολογία για το πόσο πιθανή είναι η κάθε κλάση. Μια μονάδα argmax περιλαμβάνεται επίσης σε όλες τις ακόλουθες υλοποιήσεις για να παρέχει το δείκτη της προβλεπόμενης κλάσης και περιλαμβάνεται επίσης στις εκτιμήσεις περιοχής/ισχύος.

- Όλα αυτά τα δίκτυα έχουν ακριβώς 40 κρυμμένους νευρώνες στο πρώτο επίπεδο. Αυτό δεν αντικατοπτρίζει έναν πραγματικό περιορισμό του πλαισίου. Μια πρώτη παρτίδα μοντέλων που χρησιμοποιήθηκαν για την αξιολόγηση των σχεδίων και επρόκειτο να αντικατασταθούν κάποια στιγμή, κατέληξε να παραμείνει μέχρι το τέλος. Μετά από κάποιο σημείο η αντικατάστασή τους θα απαιτούσε την επαναξιολόγηση κάθε τύπου υλοποίησης σχεδιασμού με κάθε νέο μοντέλο, προκειμένου οι συγκρίσεις αποτελεσμάτων να είναι κατατοπιστικές, γεγονός που θα απαιτούσε σημαντική χρήση του χρόνου σύνθεσης και θα σταματούσε την περαιτέρω πρόοδο για κάποιο χρονικό διάστημα.

- Κατά τη διάρκεια της σύνθεσης επιτρέπεται ένας πρακτικά απεριόριστος προϋπολογισμός χρονισμού, ώστε να μην πραγματοποιούνται βελτιστοποιήσεις χρονισμού και να μην αντισταθμίζονται με πιο σημαντικούς για το έργο στόχους.

- Η εκτίμηση της κατανάλωσης ισχύος γίνεται με προσομοίωση του συντιθέμενου κυκλώματος σε επίπεδο πύλης, αξιολογώντας 1000 δείγματα του συνόλου δεδομένων για την αναπαραγωγή ενός ρεαλιστικού περιβάλλοντος χρήσης. Η συχνότητα ρολογιού ορίζεται στην κρίσιμη συχνότητα χρονισμού του κυκλώματος που αναφέρει το εργαλείο σύνθεσης.

5.1 Γλωσσάριο συμβόλων

- N = ο αριθμός των χαρακτηριστικών εισόδου,

- M = ο αριθμός των κρυφών νευρώνων (στην περίπτωσή μας είναι πάντα 40),

- C = ο αριθμός των νευρώνων εξόδου/ αριθμός κλάσεων.

- S είναι ο αριθμός των δειγμάτων στο σύνολο δεδομένων,
- x_i είναι το i-οστό χαρακτηριστικό εισόδου,

- D^i το i-οστό δείγμα του συνόλου δεδομένων,
- D^i_j είναι η τιμή που παίρνει το j-οστό χαρακτηριστικό εισόδου στο i-οστό δείγμα του συνόλου δεδομένων,
- h_i είναι ο i-οστός κρυφός νευρώνας, που χρησιμοποιείται επίσης για να δηλώσει την τιμή εξόδου του πριν από τη δυαδικοποίηση,

- s_i είναι η έξοδος του i-οστού κρυμμένου νευρώνα μετά τη δυαδικοποίηση, οπότε $s_i = h_i \geq 0$,

- y_i είναι ο i-οστός νευρώνας εξόδου, που χρησιμοποιείται επίσης για να δηλώσει την τιμή εξόδου του,

- $W1$ = ο πίνακας βαρών του πρώτου στρώματος,

- $W2$ = ο πίνακας βαρών του δεύτερου στρώματος,

Οι σειρές αντιπροσωπεύουν νευρώνες και οι στήλες αντιπροσωπεύουν ενεργοποιήσεις εισόδου, έτσι $W1_{i,j}$ είναι το βάρος του πρώτου στρώματος που αντιστοιχεί στην σύνδεση μεταξύ του χαρακτηριστικού εισόδου x_j και του νευρώνα h_i.

6 Πλήρως συνδυαστικές πλήρως συνδεδεμένες υλοποιήσεις

Συγκρίνονται δύο αρχικές προσεγγίσεις για την υλοποίηση των πλήρως συνδεδεμένων BNNs σε ένα πλήρως συνδυαστικό μονοπάτι δεδομένων. Μόνο το πρώτο στρώμα διαφέρει μεταξύ τους, το δεύτερο στρώμα παραμένει αμετάβλητο.

6.1 Θετικό-αρνητικό άθροισμα

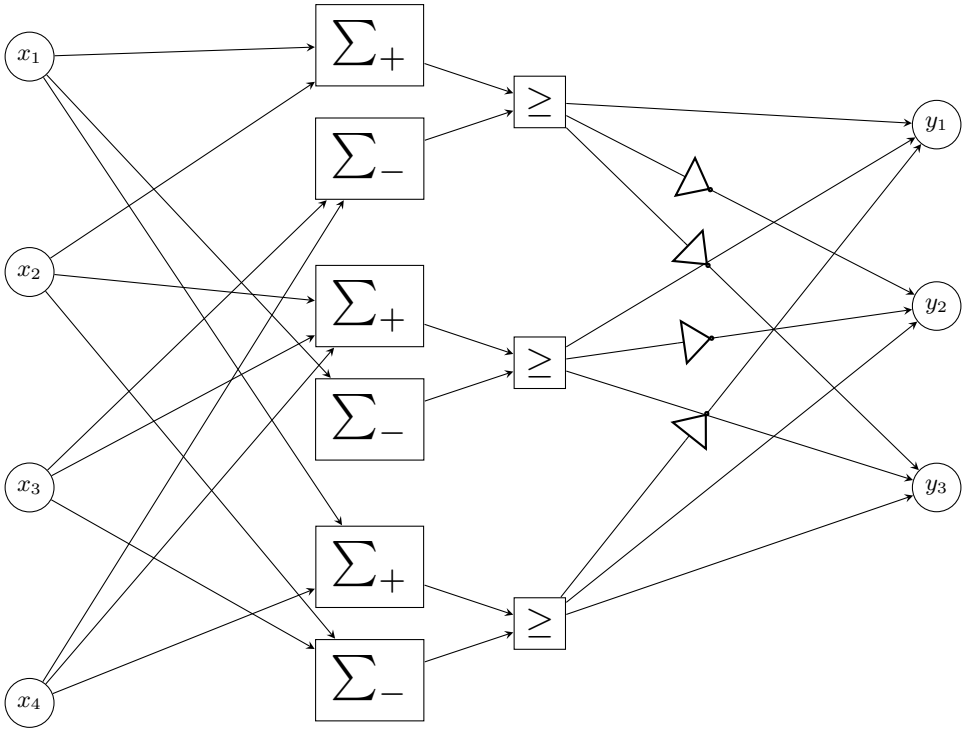

Figure 8: Υλοποίηση με διαχωρισμό θετικών και αρνητικών αθροισμάτων

Για κάθε νευρώνα στο πρώτο στρώμα υπολογίζονται δύο αθροίσματα. Σ_i^+ είναι το άθροισμα των χαρακτηριστικών εισόδου για τα οποία η σύνδεση με τον i-οστό κρυφό νευρώνα έχει θετικό βάρος , ενώ Σ_i^- το άθροισμα εκείνων που έχουν αρνητικό βάρος συνδεδεμένο. Τα δύο αθροίσματα συγκρίνονται στη συνέχεια και αν το θετικό άθροισμα είναι μεγαλύτερο ή ίσο με το αρνητικό η έξοδος του

νευρώνα είναι 1, διαφορετικά 0.

$$\Sigma_i^+ = \sum_{j=0}^{N-1} x_j[W1_{i,j} > 0]$$

$$\Sigma_i^- = \sum_{j=0}^{N-1} x_j[W1_{i,j} < 0]$$

$$h_i = \Sigma_i^+ \geq \Sigma_i^-$$

Sample code snippet:

```
assign positives[0] = + feature_array[1] + feature_array[2] +
 ↪  ... + feature_array[10];
assign negatives[0] = + feature_array[0] + feature_array[3] +
 ↪  feature_array[5];
assign hidden[0] = positives[0] >= negatives[0];
```

Η λογική πίσω από τη διάσπαση των αθροισμάτων είναι ότι η διατήρηση των λειτουργιών στη χρήση μόνο μη προσημασμένων θετικών αριθμών και η χρήση μόνο της πρόσθεσης και όχι της αφαίρεσης σημαίνει ότι απαιτούνται απλούστερες λειτουργίες και αυτό μπορεί να οδηγήσει σε μικρότερο αποτύπωμα.

Για κάθε νευρώνα του στρώματος εξόδου η τιμή του υπολογίζεται αθροίζοντας την έξοδο των κρυφών νευρώνων. Η δυαδική έξοδος του κρυμμένου νευρώνα s_j προστίθεται ως έχει στο άθροισμα του νευρώνα εξόδου y_i στην περίπτωση που το βάρος της σύνδεσής τους $W2_{i,j}$ είναι θετικό και το δυαδικό αντίστροφό του προστίθεται στο άθροισμα αν το $W2_{i,j}$ είναι αρνητικό. Αυτό ισοδυναμεί με το άθροισμα του xnor μεταξύ του διανύσματος εξόδου του κρυφού στρώματος και του διανύσματος βάρους του νευρώνα εξόδου.

$$y_i = \sum_{j=0}^{M-1} \begin{cases} s_j, & \text{if } W2_{i,j} > 0 \\ \neg s_j, & \text{if } W2_{i,j} < 0 \end{cases}$$

Code sample:

```
assign scores[0*SUM_BITS+:SUM_BITS] = + hidden_n[0] + hidden[1]
 ↪  + hidden[2] + ... + hidden_n[39];
```

6.2 Προσημασμένο άθροισμα

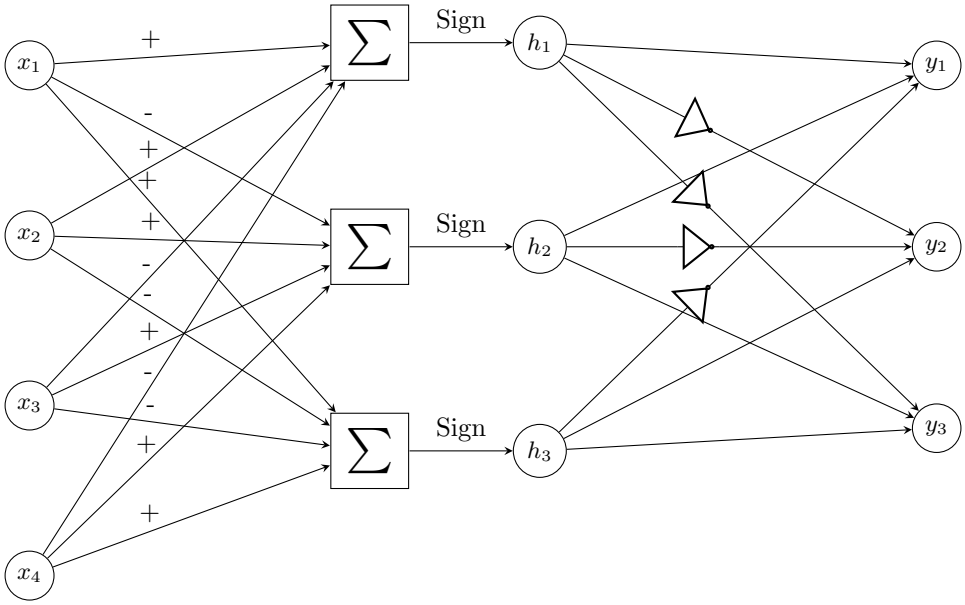

Figure 9: Εφαρμογή ενός μόνο αθροίσματος ανά νευρώνα

Σε αυτή την έκδοση υπολογίζεται ένα μόνο άθροισμα για κάθε νευρώνα. Εάν η σύνδεση μεταξύ του χαρακτηριστικού εισόδου x_j και του κρυμμένου νευρώνα h_i έχει βάρος $W1_{i,j} = 1$ προστίθεται στο άθροισμα, διαφορετικά αφαιρείται από αυτό. Ουσιαστικά η πρόσθεση του χαρακτηριστικού πολλαπλασιασμένου είτε με 1 είτε με -1 είναι hard-coded ως η προκύπτουσα πρόσθεση ή αφαίρεση αντίστοιχα. Το αποτέλεσμα στη συνέχεια συγκρίνεται με το μηδέν για να δώσει τη δυαδική έξοδο του νευρώνα. Δεδομένου ότι το αποτέλεσμα είναι ένας προσημασμένος αριθμός, αυτό σημαίνει απλώς ότι λαμβάνεται το bit του προσήμου.

$$h_i = \sum_{j=0}^{N-1} \begin{cases} +x_j, & \text{if } W1_{i,j} > 0 \\ -x_j, & \text{if } W1_{i,j} < 0 \end{cases}$$

Code sample:

```
wire signed [8:0] intra_0;
assign intra_0 = - feature_array[0] + feature_array[1] + ... +
 ↪  feature_array[10];
assign hidden[0] = intra_0 >= 0;
```

Η υλοποίηση του δεύτερου στρώματος δεν αλλάζει από τον τρόπο που είναι περιγράφεται παραπάνω.

6.2.1 Αποτελέσματα και ανάλυση

Table 2: Σύγκριση του απλού προσημασμένου αθροίσματος με την προσέγγιση διπλού μη προσημασμένου αθροίσματος

	bnnpar area(cm²)	bnnparsign area(cm²)	area change	bnnpar power(mW)	bnnparsign power(mW)	power change
Har	29.4	24.52	-16.6%	92.1	78.8	-14.4%
cardio	46.71	33.27	-28.8%	145.3	106.2	-26.9%
gasId	269.76	175.09	-35.1%	767.7	499.1	-35.0%
pendigits	42.95	33.38	-22.3%	136.8	108.9	-20.4%
winered	27.82	22.45	-19.3%	90.7	74.6	-17.8%
winewhite	26.01	20.47	-21.3%	84.6	68	-19.6%

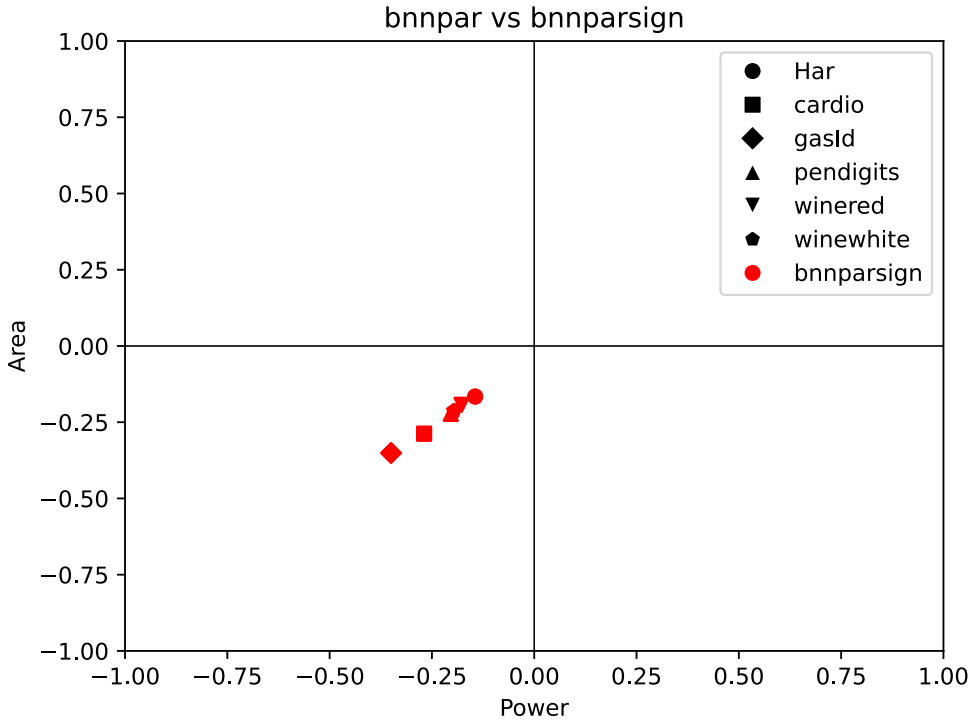

Figure 10: Σύγκριση του αθροίσματος με μονό πρόσημο με την προσέγγιση διπλών αθροισμάτων χωρίς πρόσημο

Η αρχική μου προσδοκία ήταν ότι η διάσπαση των χαρακτηριστικών σε δύο αθροίσματα για την αποφυγή αφαιρέσεων θα οδηγούσε σε καλύτερα αποτελέσματα από τη διατήρηση όλων των πράξεων για έναν νευρώνα σε μία μόνο έκφραση με βάση το σκεπτικό που συνοψίστηκε προηγουμένως. Στην πραγματικότητα αποδεικνύεται ότι η χρήση ενός ενιαίου αθροίσματος έχει απαιτήσεις σε επιφάνεια και ισχύ 20-30% χαμηλότερες από τη χρήση δύο αθροισμάτων.

Η εξήγησή μου γι' αυτό είναι ότι η διατήρηση των όρων σε ξεχωριστές εκφράσεις εμποδίζει τον μεταγλωττιστή να εντοπίζει και να μειώνει αποτελεσματικά τις κοινές υποεκφράσεις κατά τη σύνθεση. Για παράδειγμα, αν η έκφραση ενός νευρώνα περιέχει $+x_5 + x_6 - x_7$ και ένας άλλος νευρώνας περιέχει $-x_5 + x_6 + x_7$ τότε το αποτέλεσμα του $x_5 - x_7$ μπορεί να χρησιμοποιηθεί και για τους δύο νευρώνες, αλλά αν τα x_5 και x_7 δεν είναι στην ίδια έκφραση, όπως δεν θα ήταν στην υλοποίηση θετικού-αρνητικού αθροίσματος, αυτή η αριθμητική βελτιστοποίηση

δεν χρησιμοποιείται από τον μεταγλωττιστή.

6.2.2 Μείωση του ελάχιστου εύρους bit-width

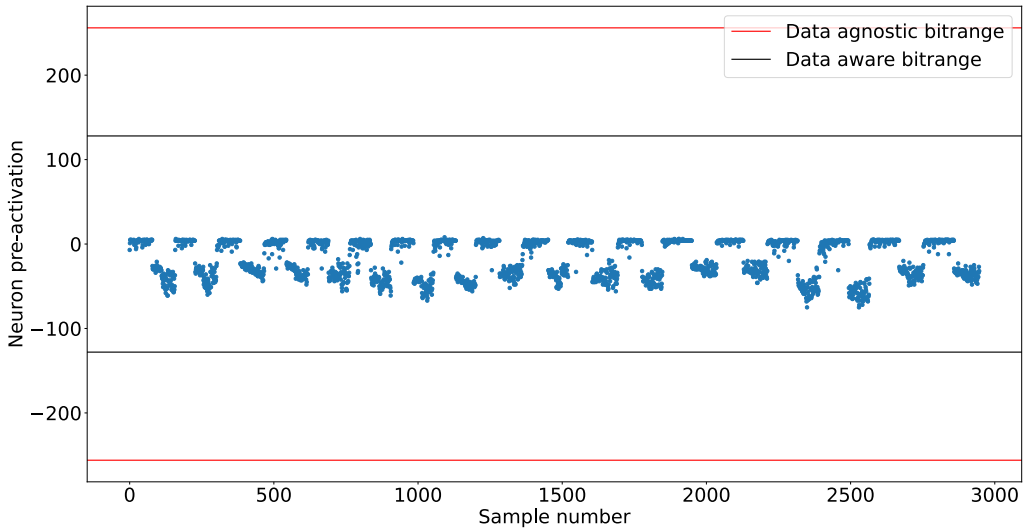

Figure 11: Τιμές προ-ενεργοποίησης για έναν κρυφό νευρώνα για κάθε δείγμα του συνόλου δεδομένων και σύγκριση του προκαθορισμένου και του ελάχιστου εύρους bit που υποστηρίζει όλες τις τιμές

Το σκεπτικό λέει ότι αν μειωθεί το εύρος bit που απαιτείται για το συνολικό άθροισμα των χαρακτηριστικών, μειώνεται και το εύρος bit που απαιτούν χαμηλότερα στο γράφημα του αθροιστή τα επιμέρους αθροίσματα από τα οποία εξαρτάται. Κατά συνέπεια, οι αθροιστές πρέπει να φιλοξενήσουν λιγότερα bits και πρέπει να υλοποιηθεί λιγότερη λογική γι' αυτούς.

Τουλάχιστον θεωρητικά, το ελάχιστο κύκλωμα για την υλοποίηση των υπολογισμών των αθροισμάτων του πρώτου στρώματος με μειωμένα πλάτη για τα αποτελέσματα θα πρέπει να είναι αυστηρά ίσο ή μικρότερο από αυτό με πλήρη πλάτη. Αυτό οφείλεται στο γεγονός ότι, δεδομένου ενός κυκλώματος που υλοποιεί τους υπολογισμούς πλήρους πλάτους, τα αποτελέσματα μειωμένου πλάτους μπορούν να ληφθούν επιλέγοντας την περιοχή bit αυτού του πλάτους από τα λιγότερο σημαντικά bits του αποτελέσματος πλήρους πλάτους. Αυτό είναι απλή επιλογή σημάτων και δεν απαιτεί πρόσθετο υλικό, επομένως η μείωση του πλάτους των νευρώνων δεν μπορεί ποτέ να απαιτήσει πρόσθετη λογική.

Μέχρι στιγμής στο πρώτο στρώμα το εύρος bit του συνολικού αθροίσματος h_i του νευρώνα έχει οριστεί να είναι αρκετά μεγάλο ώστε να χωράει οποιαδήποτε τιμή

που μπορεί να προκύψει ως αποτέλεσμα M προσθέσεων και αφαιρέσεων 4-bit μη προσημασμένων αριθμών. Η υπόθεση είναι ότι αυτό το εύρος είναι σημαντικά ευρύτερο από το εύρος των τιμών που παίρνει στην πραγματικότητα ο νευρώνας κατά την αξιολόγηση τυπικών δειγμάτων. Αυτό θα σήμαινε ότι το εύρος bit μπορεί να μειωθεί χωρίς σφάλματα λόγω υπερχείλισης ή υποχείλισης κατά την πραγματική χρήση του σχεδίου, και αυτή η μείωση θα βελτίωνε την απόδοση.

Για να ελεγχθεί αυτό υπολογίζεται το συνολικό άθροισμα κάθε νευρώνα h_i για κάθε δείγμα στο σύνολο δεδομένων. Παίρνω την ελάχιστη και τη μέγιστη τιμή αυτών των τιμών. Δεδομένου ότι όλες οι τιμές που πρέπει να πάρει το συνολικό άθροισμα περιέχονται στο εύρος μεταξύ αυτών των δύο, οι αριθμητικές πράξεις δεν χρειάζεται να εξυπηρετήσουν οποιοδήποτε εύρος μεγαλύτερο από αυτό. Έστω H_j^i η τιμή του h_i κατά την αξιολόγηση του i-οστού δείγματος του συνόλου δεδομένων και wh_i το εύρος bit του i-οστού κρυμμένου νευρώνα.

$$h_{imax} = \max_{j=0}^{S-1} H_j^i$$

$$h_{imin} = \min_{j=0}^{S-1} H_j^i$$

$$wh_i = \lceil \log_2(\max(h_{imax}, |h_{imin}| - 1)) \rceil + 1$$

6.2.3 Αποτελέσματα και ανάλυση

Table 3: Αποτελέσματα του περιορισμού του εύρους bit των νευρώνων

	bnnparsign area(cm²)	bnnparw area(cm²)	area change	bnnparsign power(mW)	bnnparw power(mW)	power change
Har	24.52	24.25	-1.1%	78.8	77.6	-1.5%
cardio	33.27	33.21	-0.2%	106.2	105.4	-0.8%
gasId	175.09	171.37	-2.1%	499.1	486.9	-2.4%
pendigits	33.38	33.97	+1.8%	108.9	109.6	+0.6%
winered	22.45	21.87	-2.6%	74.6	72.3	-3.1%
winewhite	20.47	20.36	-0.5%	68	66.7	-1.9%

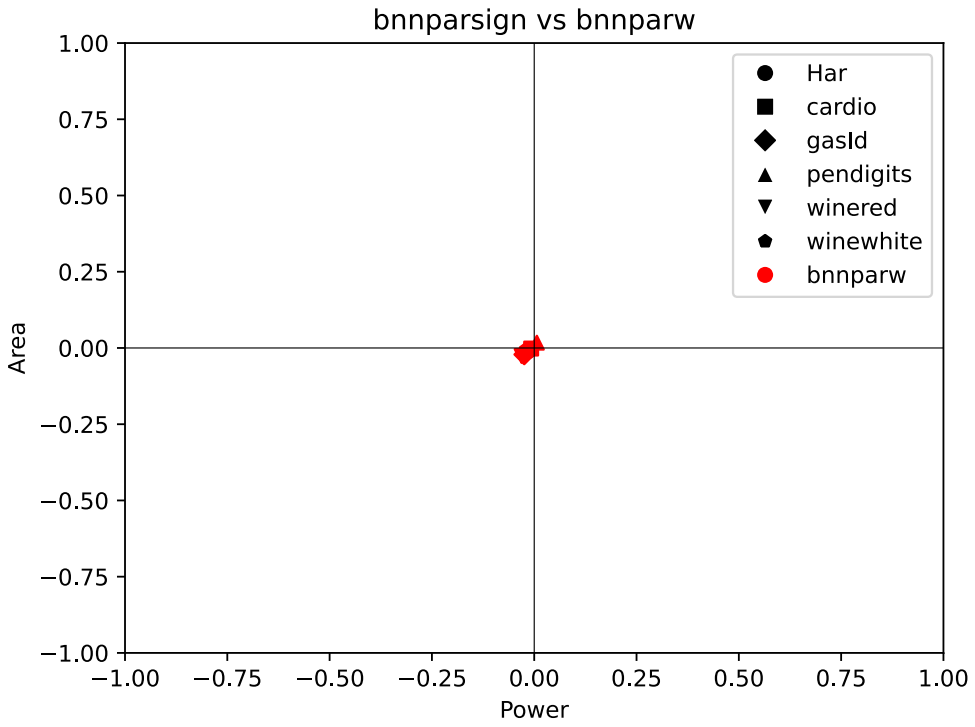

Figure 12: Αποτελέσματα του περιορισμού των bitwidths των νευρώνων

Τα αποτελέσματα είναι αμελητέα, στην περιοχή του 1-3%, και στην περίπτωση των pendigits μάλιστα επιδεινώνονται λίγο. Αυτή η επιδείνωση δεν θα έπρεπε να είναι δυνατή θεωρητικά εάν τα συνθετικά κυκλώματα είναι βέλτιστες υλοποιήσεις της περιγραφής τους. Αυτό είναι τουλάχιστον ένδειξη ότι τα αμελητέα αποτελέσματα στα άλλα σύνολα δεδομένων μπορούν να είναι καλύτερα, αν μπορέσω να τα φέρω σε μια μορφή με την οποία ο μεταγλωττιστής μπορεί να δουλέψει καλύτερα, αν και εξακολουθώ να πιστεύω ότι το κύριο πρόβλημα είναι με την ίδια την προσέγγιση. Φαίνεται ότι η περικοπή αντιβαίνει στις βέλτιστες πρακτικές και μπλοκάρει τη διαδικασία εξαγωγής μονοπατιών δεδομένων για ορισμένους νευρώνες, οπότε ορισμένες βελτιστοποιήσεις δεν εφαρμόζονται σε αυτούς και ορισμένοι που θα ήταν κοινοί πόροι δεν μοιράζονται. Δεν έχω βρει μέθοδο για να παρακάμψω αυτόν τον περιορισμό σε αυτό το σημείο.

6.2.4 Μείωση του εύρους bit των ενδιάμεσων αποτελεσμάτων

6.2.4.1 Συλλογισμός Με δεδομένο τον τελικό στόχο της εφαρμογής τεχνικών προσεγγιστικού υπολογισμού στο γράφημα αθροιστών των σχεδίων προκύπτει ένα πρόβλημα, που αναφέρθηκε προηγουμένως. Εάν η προσέγγιση, όποια και αν είναι αυτή, εφαρμόζεται ξεχωριστά στο άθροισμα κάθε νευρώνα, σχεδόν σίγουρα μπλοκάρει τη διαδικασία εξαγωγής διαδρομών δεδομένων από την εκτέλεση αριθμητικών βελτιστοποιήσεων, όπως η μείωση και ο διαμοιρασμός κοινών υποεκφράσεων μεταξύ των νευρώνων. Το αποτέλεσμα είναι M ξεχωριστά δέντρα προσεγγιστικού αθροιστή. Ακόμη και με δεδομένο ότι η λογική μείωση που κερδίζεται από τις προσεγγιστικές προσθέσεις για κάθε δέντρο αθροιστή είναι υπεραρκετή για να αντισταθμίσει το χαμένο όφελος από την κοινή χρήση ενδιάμεσων αποτελεσμάτων, αυτή μπορεί να είναι μια περιττή παραχώρηση.

Για να εκτιμήσω πόσο μεγάλη μπορεί να είναι η αρνητική επίδραση μιας τεχνικής προσέγγισης που δεν λαμβάνει υπόψη την κοινή χρήση μεταξύ των νευρώνων, μειώνω τα πλάτη bit των ενδιάμεσων αποτελεσμάτων του αθροίσματος κάθε νευρώνα. Παρόλο που αυτό θα βοηθούσε εύλογα για έναν μεμονωμένο νευρώνα, αναμένω ότι θα προκαλέσει τη διακοπή του διαμοιρασμού πόρων. Δεδομένου του πόσο μεγάλη είναι η αρνητική επίδραση, μπορώ να ελέγξω αν αυτό είναι ένα πρόβλημα που θα έπρεπε να διορθωθεί πριν η προσέγγιση μπορεί να εφαρμοστεί με σιγουριά.

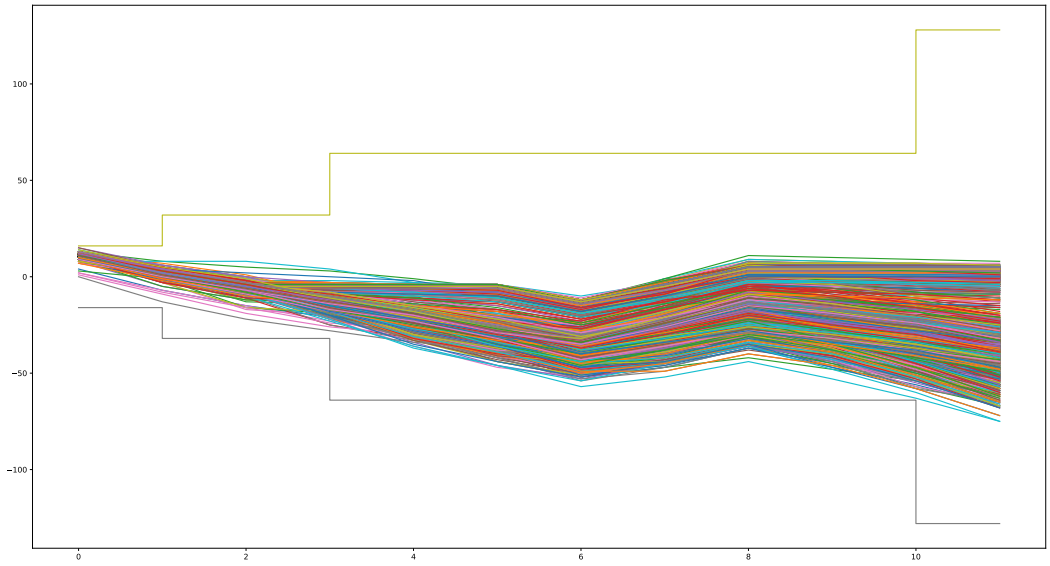

Figure 13: Τα ενδιάμεσα αθροίσματα του διαδοχικού υπολογισμού του κρυμμένου νευρώνα 28 για όλο το σύνολο δεδομένων Har και το εύρος των bit που απαιτείται για την υποστήριξη κάθε βήματος

6.2.4.2 Εφαρμογή Το σωρευτικό άθροισμα κατά μήκος των χαρακτηριστικών εισόδου πολλαπλασιασμένων με το βάρος της σύνδεσής τους με τον νευρώνα i υπολογίζεται για κάθε κρυφό νευρώνα και για κάθε δείγμα του συνόλου δεδομένων. Στη συνέχεια, υπολογίζεται η μέγιστη και η ελάχιστη τιμή σε όλα τα δείγματα σε κάθε βήμα του σωρευτικού αθροίσματος για έναν συγκεκριμένο νευρώνα. Με παρόμοιο τρόπο με τη μέθοδο που περιγράφηκε παραπάνω για τη μείωση του εύρους bit του συνολικού αποτελέσματος των πράξεων του νευρώνα, οι πράξεις γράφονται διαδοχικά με το αποτέλεσμα της καθεμιάς να έχει το εύρος bit που ορίζεται με βάση το εύρος τιμών για το ισοδύναμο βήμα του αθροίσματος για όλο το σύνολο δεδομένων.

Έστω $h_{i,j}$ το αποτέλεσμα της τιμής προ-ενεργοποίησης του i-οστού κρυμμένου νευρώνα λαμβάνοντας υπόψη μόνο τα χαρακτηριστικά εισόδου x_0 έως x_j, ή ισοδύναμα η τιμή του h_i αν τα x_{j+1} έως x_{N-1} είναι καλυμμένα με μηδέν.

$$h_{i,j} = \sum_{k=0}^{j} x_k W1_{i,k}$$

$$hmax_{i,j} = \max_{l=0}^{S-1} \sum_{k=0}^{j} D_k^l W1_{i,k}$$

$$hmin_{i,j} = \min_{l=0}^{S-1} \sum_{k=0}^{j} D_k^l W1_{i,k}$$

$$wh_{i,j} = \lceil \log_2(\max(hmax_{i,j}, |hmin_{i,j}| - 1)) \rceil + 1$$

Μερικές φορές, λόγω της σειράς των προσθέσεων και αφαιρέσεων, το πλάτος που απαιτείται σε ένα μεταγενέστερο βήμα είναι μικρότερο από αυτό ενός προηγούμενου βήματος. Αυτό οφείλεται στο γεγονός ότι κάθε δείγμα για το οποίο θα υπήρχε υπερχείλιση στο προηγούμενο βήμα με το μικρότερο πλάτος θα υποχείλιζε σε κάποιο επόμενο χαρακτηριστικό πίσω στο εύρος που υποστηρίζει. Αυτό έχει ληφθεί υπόψη. Εάν ένα πλάτος του αποτελέσματος μιας επόμενης λειτουργίας είναι μικρότερο, το πλάτος bit της προηγούμενης πρόσθεσης/αφαίρεσης απλά τίθεται σε αυτή τη μικρότερη τιμή.

$$wh'_{i,j} = \min_{k=j}^{N-1} wh_{i,j}$$

Διαδοχικές πράξεις που έχουν το ίδιο εύρος bit στο αποτέλεσμα ομαδοποιούνται και εκφράζονται σε verilog ως ένα ενιαίο άθροισμα. Δεν έχω επιβεβαιώσει αν αυτό όντως επηρεάζει καθόλου το αποτέλεσμα της σύνθεσης, αλλά φαίνεται να συμμορφώνεται περισσότερο με τις συστάσεις του οδηγού βέλτιστων πρακτικών.

6.2.4.3 Αποτελέσματα και ανάλυση

Table 4: Αποτελέσματα της συρρίκνωσης των ενδιάμεσων αποτελεσμάτων

	bnnparw area(cm²)	bnnparstepw area(cm²)	area change	bnnparw power(mW)	bnnparstepw power(mW)	power change
Har	24.25	23.16	-4.5%	77.6	73.7	-5.0%
cardio	33.21	39.29	+18.3%	105.4	125.5	+19.1%
gasId	171.37	326.98	+90.8%	486.9	935.2	+92.1%
pendigits	33.97	37.09	+9.2%	109.6	120.3	+9.8%
winered	21.87	22.78	+4.2%	72.3	75.3	+4.1%
winewhite	20.36	20	-1.8%	66.7	65.8	-1.3%

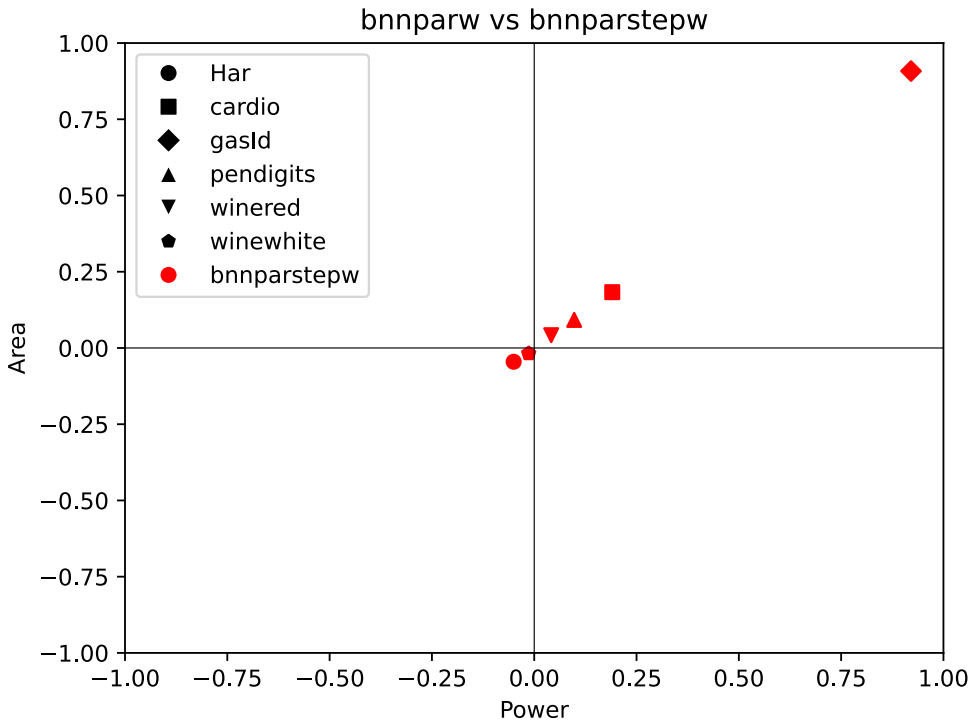

Figure 14: Αποτελέσματα της συρρίκνωσης ενδιάμεσων αποτελεσμάτων

Τα αποτελέσματα ήταν θετικά για δύο από τα μοντέλα με τον μικρότερο αριθμό χαρακτηριστικών εισόδου N, πράγμα που σημαίνει ότι η μέθοδος αυτή απέδωσε καλύτερα από την αρχική προσδοκία. Ωστόσο, η απώλεια βελτιστοποίησης από τη μείωση των κοινών λειτουργιών παρουσιάζει σαφή κλιμάκωση με τον αριθμό N των χαρακτηριστικών εισόδου. Με περισσότερα στοιχεία προς άθροιση, θα έπρεπε να προκύπτουν περισσότερες κοινές υποεκφράσεις προς βελτιστοποίηση, οπότε αυτό ανταποκρίνεται στις προσδοκίες.

Στο δίκτυο με το μεγαλύτερο N, αυτό που ανήκει στο gasId, το οποίο έχει 128 χαρακτηριστικά αισθητήρων, η κατάσταση έχει γίνει αρκετά σοβαρή ώστε να διπλασιαστούν σχεδόν οι απαιτήσεις σε επιφάνεια και ισχύ. Αυτό δείχνει ότι το πρόβλημα είναι πράγματι σημαντικό όταν υπάρχουν περισσότερες από ένα μικρό αριθμό εισόδων και θα πρέπει να αναζητηθούν τρόποι για την αντιμετώπισή του.

6.2.5 Προκαταβολική αριθμητική βελτιστοποίηση

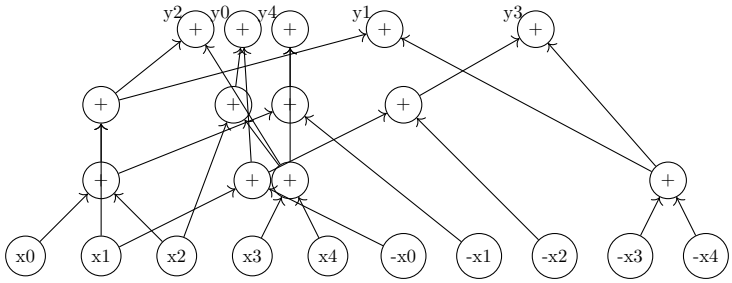

Figure 15: Υλοποίηση ενός δυαδικού στρώματος βάρους με προκαθορισμένη σειρά προσθέσεων

6.2.5.1 Αιτιολόγηση Με βάση τα αποτελέσματα από τα παραπάνω επιχειρώ να πάρω την αριθμητική έκφραση που υπολογίζει το σχέδιο μετά τη σύνθεση, με σκοπό να προσαρμόσω σε αυτήν μεταγενέστερες τεχνικές προσέγγισης, αντί να υπαγορεύουν οι προσεγγίσεις το γράφημα των πράξεων και να χάνονται αυτά τα πλεονεκτήματα.

Ο μεταγλωττιστής σχεδίασης παρέχει μια αναφορά "*resource and datapath extraction*", η οποία, παραθέτοντας τον οδηγό χρήσης, "αναλύει το αριθμητικό περιεχόμενο της σχεδίασης και παρέχει ανατροφοδότηση, ώστε να μπορείτε να βελτιώσετε τον κώδικα RTL όπως απαιτείται". Σε αυτή την αναφορά οι αριθμητικές πράξεις που εκτελούνται μετά τη βελτιστοποίηση από κάθε μπλοκ datapath περιγράφονται στην ενότητα πόρων του μπλοκ. Από αυτό θα μπορούσε να ανακατασκευαστεί ένα γράφημα πρόσθεσης/αφαίρεσης από τα στοιχεία εισόδου στις εξόδους του στρώματος με σχετικά απλή ανάλυση.

Δυστυχώς, η έκθεση δεν παρέχει αντιστοίχιση μεταξύ των συμβόλων που χρησιμοποιεί για τις μεταβλητές εισόδου και εξόδου των μπλοκ datapath και των αντίστοιχων σημάτων στην αρχική σχεδίαση. Εξαιτίας αυτού, ο ανακατασκευασμένος γράφος αθροιστή δεν μπορεί να χρησιμοποιηθεί για την υλοποίηση των επιπέδων του δικτύου, πριν οι είσοδοι και οι έξοδοι επισημανθούν με άλλο τρόπο.

Ως εναλλακτική λύση, αναζητούνται στη βιβλιογραφία αλγόριθμοι ή ευρετικές μέθοδοι που θα εκτελούσαν μια αντίστοιχη αριθμητική βελτιστοποίηση με αυτή που παρέχει ο Design Compiler. Δεν φαντάζει πολύ αισιόδοξο ότι υπάρχει μια de facto πρότυπη μέθοδος για τέτοιες περιπτώσεις και ίσως μάλιστα να είναι αυτή που χρησιμοποιείται κάτω από την κουκούλα του μεταγλωττιστή, οπότε οι λειτουργίες που βρίσκονται είναι ακριβώς ή σχεδόν οι ίδιες.

Var	Type	Data Class	Width	
I1	PI	Unsigned	4	
I2	PI	Unsigned	4	
I3	PI	Unsigned	4	
I4	PI	Unsigned	4	
I5	PI	Unsigned	4	
I6	PI	Unsigned	4	
I7	PI	Unsigned	4	
I8	PI	Unsigned	4	
I9	PI	Unsigned	4	
I10	PI	Unsigned	4	
I11	PI	Unsigned	4	
I12	PI	Unsigned	8	
I13	PI	Unsigned	7	
I14	PI	Unsigned	8	
T1540	IFO	Unsigned	5	I5 + I6 (winered_bnn1_bnnpar.v:44winered_bnn1_bnnpar.v:48winered_bnn1_bnnpar.v
T1848	IFO	Unsigned	6	I4 + T1540 (winered_bnn1_bnnpar.v:80winered_bnn1_bnnpar.v:99winered_bnn1_bnn
T1994	IFO	Unsigned	5	I11 + I8 (winered_bnn1_bnnpar.v:48winered_bnn1_bnnpar.v:51winered_bnn1_bnnpar
T1567	IFO	Unsigned	5	I10 + I7 (winered_bnn1_bnnpar.v:40winered_bnn1_bnnpar.v:44winered_bnn1_bnnpar.v:48wine
T1548	IFO	Unsigned	6	I9 + T1567 (winered_bnn1_bnnpar.v:40winered_bnn1_bnnpar.v:44winered_bnn1_bnn
T2045	IFO	Unsigned	7	T1994 + T1548 (winered_bnn1_bnnpar.v:48winered_bnn1_bnnpar.v:75winered_bnn1_
T2042	IFO	Unsigned	8	I3 + T1848 + T2045 (winered_bnn1_bnnpar.v:99)
T158	IFO	Unsigned	5	I1 + I2 (winered_bnn1_bnnpar.v:39winered_bnn1_bnnpar.v:91winered_bnn1_bnnpar.v
T1388	IFO	Unsigned	5	I2 + I4 (winered_bnn1_bnnpar.v:43winered_bnn1_bnnpar.v:47winered_bnn1_bnnpar.v
T1690	IFO	Unsigned	6	I3 + T1388 (winered_bnn1_bnnpar.v:64winered_bnn1_bnnpar.v:83winered_bnn1_bnn
T1932	IFO	Unsigned	7	T1540 + T1690 (winered_bnn1_bnnpar.v:179winered_bnn1_bnnpar.v:183)
T1928	IFO	Unsigned	8	T1548 + T1932 (winered_bnn1_bnnpar.v:183)
T1954	IFO	Unsigned	6	I1 + T1994 (winered_bnn1_bnnpar.v:63winered_bnn1_bnnpar.v:184)
T1333	IFO	Unsigned	6	I8 + T1388 (winered_bnn1_bnnpar.v:43)
T1609	IFO	Unsigned	7	I11 + T1548 (winered_bnn1_bnnpar.v:40winered_bnn1_bnnpar.v:44winered_bnn1_bn
T98	IFO	Unsigned	5	I1 + I3 (winered_bnn1_bnnpar.v:44winered_bnn1_bnnpar.v:52winered_bnn1_bnnpar.v
T1950	IFO	Unsigned	6	T98 + T1540 (winered_bnn1_bnnpar.v:44winered_bnn1_bnnpar.v:135)
T1948	IFO	Unsigned	8	T1609 + T1950 (winered_bnn1_bnnpar.v:44)
T1946	IFO	Unsigned	8	I8 + T1950 (winered_bnn1_bnnpar.v:135)
T1595	IFO	Unsigned	8	T1388 + T1609 (winered_bnn1_bnnpar.v:136)
T1944	IFO	Unsigned	6	T1388 + T1540 (winered_bnn1_bnnpar.v:68winered_bnn1_bnnpar.v:75)
T2040	IFO	Unsigned	8	T1944 + T2045 (winered_bnn1_bnnpar.v:75)
T1714	IFO	Unsigned	7	I8 + T1548 (winered_bnn1_bnnpar.v:67winered_bnn1_bnnpar.v:107winered_bnn1_bn
T1704	IFO	Unsigned	8	T98 + T1714 (winered_bnn1_bnnpar.v:67)
T1940	IFO	Unsigned	8	I11 + T1944 (winered_bnn1_bnnpar.v:68)

Figure 16: Ένα παράδειγμα της αναφοράς εξαγωγής ενός μπλοκ datapath. Επισημαίνεται μια ενδιάμεση τιμή που επαναχρησιμοποιείται πολλές φορές.

6.2.5.2 Εφαρμογή Μια διατύπωση του προβλήματος είναι η ακόλουθη: Δεδομένου ενός καταλόγου εκφράσεων της γενικής μορφής $y_i = x_0 + x_1 - x_2 - \ldots + x_n$ στην οποία οι τελεστές μπορούν να διαμοιραστούν μεταξύ των εκφράσεων

βρείτε τον ελάχιστο αριθμό προσθέσεων ή αφαιρέσεων που πρέπει να εκτελεστούν για την αξιολόγηση όλων των εκφράσεων.

Το πρόβλημα αποδεικνύεται σημαντικά λιγότερο καλά μελετημένο από ό,τι αρχικά αναμενόταν. Ενώ η παραπλανητικά απλή περιγραφή υποδηλώνει έναν απλό τρόπο για την απάντησή του, είναι NP-Complete δύσκολο, πιο συγκεκριμένα στην οικογένεια MaxSNP των προβλημάτων βελτιστοποίησης. Ως άμεσο αποτέλεσμα επιχειρούνται μόνο προσεγγιστικές λύσεις. Ο [32] αναζητά ακριβείς λύσεις αξιοποιώντας επιλυτές SAT, αλλά καταφέρνει να το κάνει να λειτουργήσει μόνο για πολύ μικρά μεγέθη πινάκων μέχρι 8 x 8. Δεν βρέθηκαν πολλά άλλα για το ακριβές σενάριο παραπάνω, αλλά ένα αρκετά κοντινό πρόβλημα που έχει να κάνει με την παραγοντοποίηση παρόμοιων λιστών εκφράσεων της μορφής $y_i = x_0 \oplus x_1 \oplus ... \oplus x_n$ με τη χρήση των ελάχιστων δυνατών πράξεων XOR εξετάζεται ενεργά χάρη σε ορισμένες εφαρμογές στον τομέα των κρυπτογραφικών επιταχυντών. Και τα δύο ανήκουν στην οικογένεια προβλημάτων συντομότερου γραμμικού προγράμματος.

Επιλέγω να δοκιμάσω να χρησιμοποιήσω πρώτα τον αλγόριθμο παραγοντοποίησης του Paar [33]. Είναι παλαιότερος από τις περισσότερες ευρετικές μεθόδους που έχουν εφαρμοστεί στο πρόβλημα της παραγοντοποίησης XOR, αλλά έχει το πλεονέκτημα ότι δεν εκμεταλλεύεται την ακύρωση όρων. Χάρη στην ιδιότητα $x \oplus x = 0$ ορισμένες βέλτιστες λύσεις στο πρόβλημα XOR περιλαμβάνουν δύο υποεκφράσεις που περιέχουν τον ίδιο όρο x και συνδυάζονται με XOR για να παράγουν μια επιθυμητή έκφραση που δεν περιλαμβάνει τον όρο x. Οι ευρετικές μέθοδοι που αναπτύχθηκαν μετά τον Paar εκμεταλλεύονται αυτό το χαρακτηριστικό, και ενώ υπάρχει ένας παραλληλισμός μεταξύ αυτού και της ακύρωσης των όρων των αντιθέτων στο σενάριό μας ($x - x = 0$ ή $x + -x = 0$) δεν έχω καταφέρει να βρω τις προσαρμογές που απαιτούνται για να εφαρμόσω τις ιδέες τους στο νέο πεδίο. Έτσι, δίνω προτεραιότητα στην πιο απλή μέθοδο, που μεταφράζεται άμεσα στη χρήση της πρόσθεσης στη θέση της ισοτιμίας.

Δοκιμάζω επίσης μια μικρή τροποποίηση της αρχικής διαδικασίας ώστε να είναι άμεσα συμβατή με εκφράσεις που περιλαμβάνουν αφαιρέσεις.

6.2.5.3 Αποτελέσματα και ανάλυση

Table 5: Επίδραση της προκαταβολικής αριθμητικής βελτιστοποίησης με την ευρετική του Paar

	bnnparsign area(cm²)	bnnpaar area(cm²)	area change	bnnparsign power(mW)	bnnpaar power(mW)	power change
Har	24.52	17.42	-29.0%	78.8	57.2	-27.4%
cardio	33.27	38.74	+16.4%	106.2	124.1	+16.9%
gasId	175.09	281.55	+60.8%	499.1	807.6	+61.8%
pendigits	33.38	35.43	+6.1%	108.9	114.6	+5.2%
winered	22.45	18.55	-17.4%	74.6	62.6	-16.1%
winewhite	20.47	18.01	-12.0%	68	59.8	-12.1%

Table 6: Σύγκριση του τροποποιημένου ευρετικού συστήματος Paar με το αρχικό

	bnnpaar area(cm²)	bnnpaarter area(cm²)	area change	bnnpaar power(mW)	bnnpaarter power(mW)	power change
Har	17.42	18.73	+7.5%	57.2	60.8	+6.3%
cardio	38.74	35.97	-7.2%	124.1	116	-6.5%
gasId	281.55	261.38	-7.2%	807.6	759.7	-5.9%
pendigits	35.43	32.22	-9.1%	114.6	107.3	-6.4%
winered	18.55	17.47	-5.8%	62.6	59.6	-4.8%
winewhite	18.01	16.65	-7.6%	59.8	55.9	-6.5%

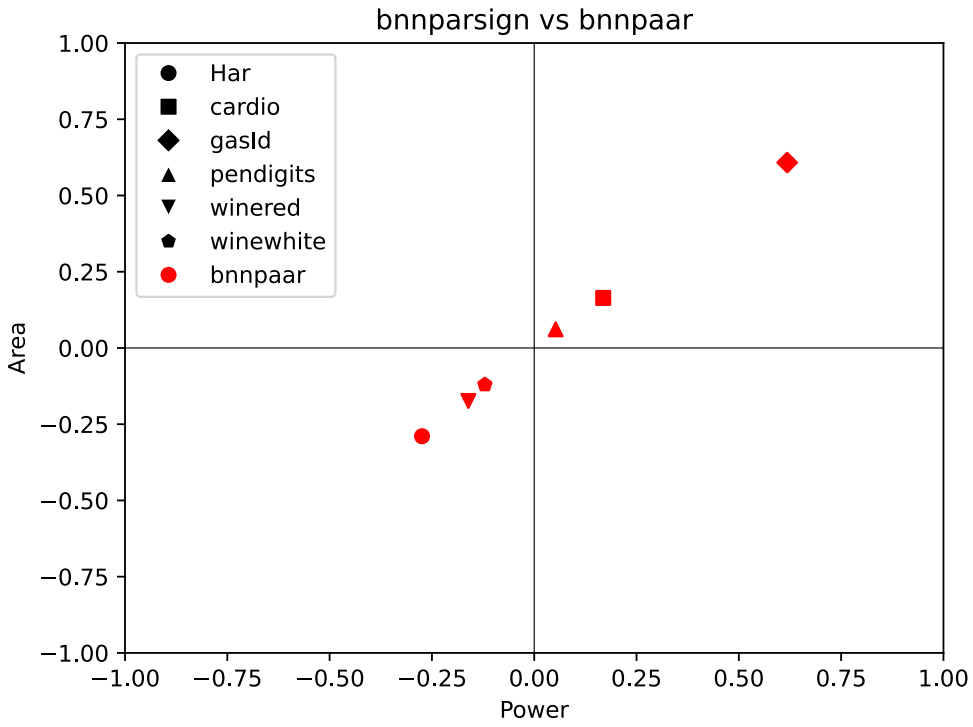

Figure 17: Επίδραση της προκαταβολικής αριθμητικής βελτιστοποίησης με την ευρετική του Paar

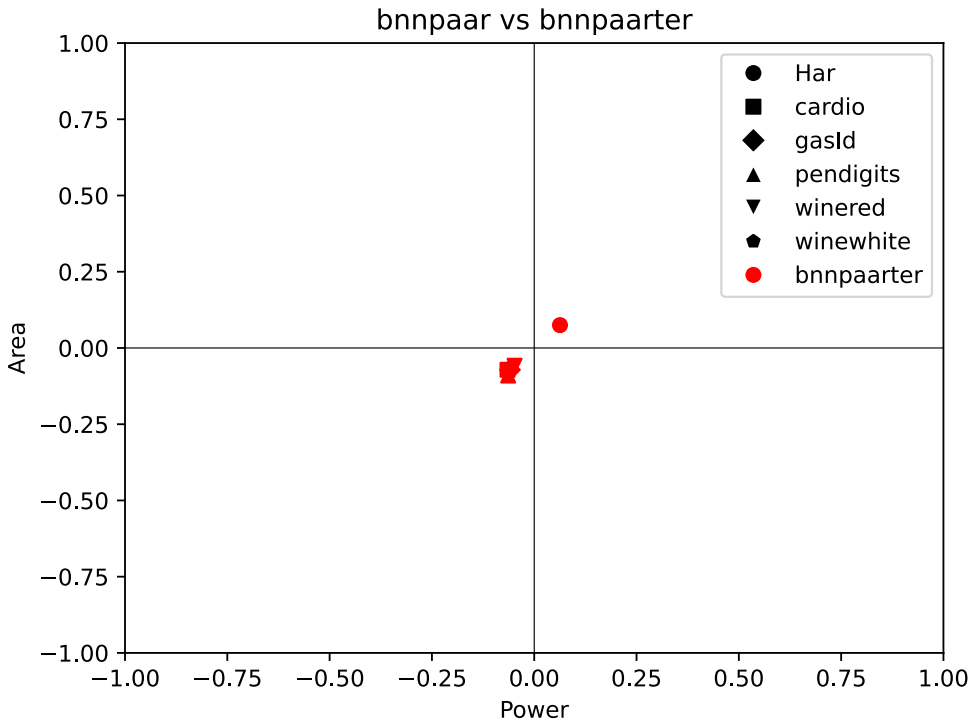

Figure 18: Σύγκριση της τροποποιημένης ευρετικής του Paar με την αρχική

Περίμενα ότι είτε:

1. Τα αποτελέσματα της σκληρής αντιστοίχισης της σειράς των πράξεων για τον υπολογισμό του αποτελέσματος της προ-ενεργοποίησης των νευρώνων με τη χρήση της ευρετικής του Paar θα είναι σημαντικά χειρότερα από το αποτέλεσμα που θα είχαμε αν αφήναμε τον Design Compiler να χρησιμοποιήσει τα αποτελέσματα της δικής του ευρετικής βελτιστοποίησης, αφού θα έπρεπε να εφαρμόζει τις καλύτερες διαθέσιμες. Σε αυτή την περίπτωση η προσπάθεια μιας εναλλακτικής λύσης για την εκ των προτέρων βελτιστοποίηση των λειτουργιών είναι πιθανότατα σπατάλη κόπου, επειδή η εύρεση μιας ανταγωνιστικής ευρετικής θα ήταν δυσκολότερη από την ανάλυση των αποτελεσμάτων της δικής τους λύσης.

2. Τα αποτελέσματα θα είχαν αμελητέα διαφορά, επειδή οι ευριστικές που χρησιμοποιούνται είναι συναφείς και/ή η ποιότητα των αποτελεσμάτων που μπορεί να αναμένεται από τις τρέχουσες μεθόδους για λογικούς

προϋπολογισμούς compute χτυπά ένα ορισμένο ανώτατο όριο για τις διάφορες προσεγγίσεις. Στην περίπτωση αυτή η εφαρμογή τεχνικών προσέγγισης στο εκτιμώμενο γράφημα αριθμητικών πράξεων μπορεί να προχωρήσει.

Τα αποτελέσματα δείχνουν ότι, αν και δεν είναι συνεπής σε όλα τα δίκτυα, υπάρχει βελτίωση της τάξης του 20-30% στις εκτιμήσεις εμβαδού και ισχύος των μικρότερων δικτύων.

Δυστυχώς, το ζήτημα που προσπαθώ να αντιμετωπίσω έχει να κάνει με τις απώλειες απόδοσης από τη διακοπή των αριθμητικών βελτιστοποιήσεων του μεταγλωττιστή που κλιμακώνονται με το μέγεθος του μοντέλου, και η σχετική απόδοση της εναλλακτικής ευρετικής κλιμακώνεται αντιστρόφως ανάλογα με το μέγεθος αυτό. Αυτό σημαίνει ότι η επιχειρούμενη διόρθωση δεν μπορεί να εφαρμοστεί στις περιπτώσεις που τη χρειάζονται περισσότερο, οπότε το υποκείμενο πρόβλημα παραμένει άλυτο.

Η τριαδική εκδοχή της ευρετικής του Paar υπερέχει της αρχικής με σχετικά σταθερή αναλογία, με εξαίρεση το δίκτυο του μοντέλου Har. Αυτό μου δίνει κάποια ελπίδα ότι η εφαρμογή πιο προηγμένων ευρετικών που χρησιμοποιούνται στα συντομότερα γραμμικά προγράμματα τροποποιημένα για τη συγκεκριμένη περίπτωση χρήσης θα αυξήσει το όριο μεγέθους του δικτύου για το οποίο μπορούν να βελτιωθούν τα αποτελέσματα.

7 Ακολουθιακή εκτέλεση

7.1 Συλλογισμός

Οι ταξινομήσεις που εξετάζονται εδώ δεν είναι ως επί το πλείστον χρονικά κρίσιμες και δεν απαιτούν υψηλή απόδοση. Η αξιολόγηση της ποιότητας ενός κρασιού κάθε δευτερόλεπτο ή πολλαπλές φορές ανά δευτερόλεπτο δεν προσφέρει μεγαλύτερη αξία από ό,τι η αξιολόγηση κάθε λίγα λεπτά. Δεδομένου ότι ο χρόνος είναι ο λιγότερο πολύτιμος πόρος για τους σκοπούς μας, μπορεί να ανταλλαχθεί με τη μείωση των απαιτήσεων σε επιφάνεια και ισχύ.

Σε αυτή την προσέγγιση ένα δέντρο μονής πρόσθεσης θα υπολογίσει τη συνολική τιμή ενός νευρώνα h_i αθροίζοντας τα σταθμισμένα χαρακτηριστικά σε έναν μόνο κύκλο και θα εκτελέσει το ίδιο για έναν διαφορετικό νευρώνα του ίδιου στρώματος στον επόμενο κύκλο.

Με αυτόν τον τρόπο το γράφημα αθροιστή που περιλαμβάνει τους υπολογισμούς που απαιτούνται για ολόκληρο το στρώμα στην πλήρως παράλληλη έκδοση μειώνεται στο δέντρο αθροιστή που απαιτεί η εξαγωγή συμπερασμάτων ενός μόνο νευρώνα. Αυτό μπορεί να αντισταθμίσει με το παραπάνω την κωδικοποίηση της κατάλληλης στάθμισης των χαρακτηριστικών πριν από τη συσσώρευσή τους από ό,τι πρέπει να γίνει τώρα, δεδομένου ότι οι πράξεις δεν είναι απλώς hard-coded στο κύκλωμα.

7.2 Υλοποίηση

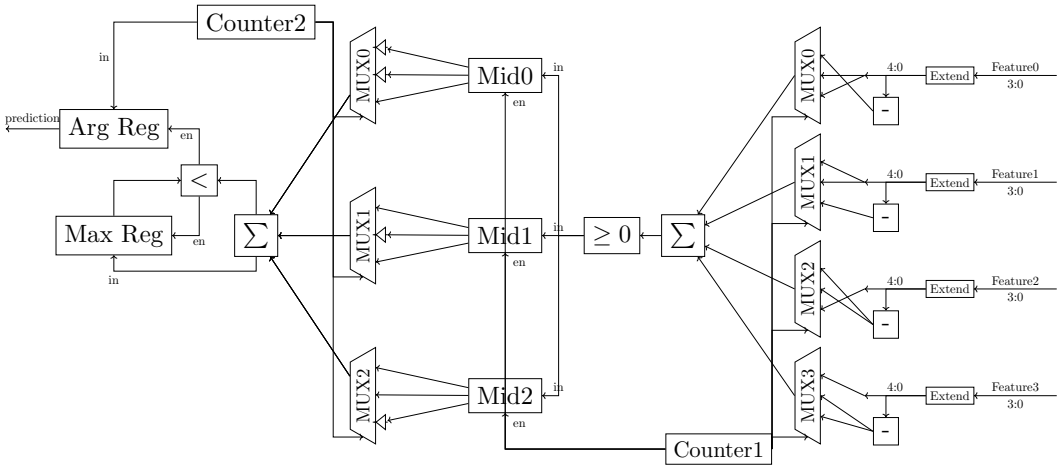

Figure 19: Μια διαδοχική υλοποίηση BNN με ένα μόνο καλώδιο

Ένας πολυπλέκτης αντιστοιχεί σε ένα συγκεκριμένο χαρακτηριστικό εισόδου και επιλέγει ποια "στάθμιση" αυτού του χαρακτηριστικού χρειάζεται ο εκάστοτε εξεταζόμενος νευρώνας.

Ας ονομάσουμε την τιμή που επιλέγεται να αντιπροσωπεύει το j-οστό χαρακτηριστικό εισόδου στον i-οστό κύκλο αξιολόγησης του επιπέδου $x_{i,j}$. Ο τρόπος με τον οποίο προκύπτει από την αρχική είσοδο x_j είναι όπως αναφέρθηκε προηγουμένως:

$$x_{i,j} = \begin{cases} x_j, & \text{if } W1_{i,j} = 1 \\ -x_j, & \text{if } W1_{i,j} = -1 \end{cases}$$

Οι επιλεγμένες σταθμισμένες είσοδοι επεξεργάζονται στη συνέχεια από ένα γενικό αθροιστικό δέντρο N εισόδων που παρέχει το πρόσημο του αθροίσματός τους, στον i-οστό κύκλο $s_i = sign(\sum_{j=0}^{N-1} x_{i,j})$. Μόνο το bit του προσήμου πρέπει να αποθηκευτεί για να είναι διαθέσιμο για τους επόμενους υπολογισμούς από το επόμενο επίπεδο, η τιμή προ-ενεργοποίησης h_i μπορεί να απορριφθεί με ασφάλεια. Έτσι, ένας καταχωρητής 1 bit δεικτοδοτείται από τον μετρητή για να αποθηκεύσει την έξοδο του τρέχοντος νευρώνα. Αυτή η διαδικασία απαιτεί τόσους κύκλους για την αξιολόγηση ενός στρώματος όσοι και ο αριθμός των νευρώνων του, άρα M κύκλους ρολογιού για το πρώτο στρώμα.

Ένα σύστημα ανιχνεύει πότε ο τελευταίος νευρώνας του στρώματος έχει ολοκληρώσει τις λειτουργίες και δίνει τη σκυτάλη στο επόμενο στρώμα για να ξεκινήσει. Το επόμενο στρώμα αντιστρέφει υπό όρους τα χαρακτηριστικά εισόδου που λαμβάνει πριν τα περάσει στις γραμμές δεδομένων των πολυπλεκτών στη λογική που το πρώτο στρώμα τα αντιστρέφει.

Στο δεύτερο και τελευταίο στρώμα τα αποτελέσματα δεν χρειάζεται να αποθηκευτούν καθόλου. Επειδή η έξοδος των νευρώνων υπολογίζεται μία προς μία, η λειτουργία argmax μπορεί να ενσωματωθεί σε αυτή τη διαδικασία. Χρησιμοποιούνται δύο καταχωρητές, ο ένας κρατάει τη μεγαλύτερη έξοδο ενός νευρώνα που έχει δει μέχρι στιγμής y_{max} και ο άλλος τον δείκτη του προαναφερθέντος νευρώνα στο στρώμα, ο οποίος λαμβάνεται από την τιμή του μετρητή κύκλων του στρώματος.

Εάν το αποτέλεσμα της πράξης popcount στις τρέχουσες σταθμισμένες εισόδους στον i-οστό κύκλο της δραστηριότητας y_i του δεύτερου στρώματος είναι μεγαλύτερο από το προηγούμενο καλύτερο y_{max}, το νέο αποτέλεσμα αποθηκεύεται ως το νέο καλύτερο μέχρι στιγμής και ο δείκτης αντικαθίσταται από τον αριθμό κύκλων i. Πέρα από την εξοικονόμηση πολλών flip-flops αφαιρείται η επιβάρυνση

της πρόσθετης μονάδας argmax που έπρεπε να ενεργοποιηθεί μετά το δεύτερο στρώμα.

Το δεύτερο στρώμα παίρνει τόσους κύκλους όσες και οι κλάσεις που πρέπει να εξεταστούν για πρόβλεψη. Συνεπώς, η πλήρης εξαγωγή συμπερασμάτων καταλαμβάνει $M + C$ κύκλους ρολογιού. Και πάλι πρέπει να δοθεί ένα σήμα επαναφοράς μεταξύ διαδοχικών εκτελέσεων.

7.2.1 Αποτελέσματα και ανάλυση

Table 7: Σύγκριση ακολουθιακών σχεδιασμών ενιαίου δένδρου πρόσθεσης με ισοδύναμους συνδυαστικούς σχεδιασμούς.

	bnnparw area(cm²)	bnnrolx area(cm²)	area change	bnnparw power(mW)	bnnrolx power(mW)	power change
Har	24.25	9.14	-62.3%	77.6	39	-49.7%
cardio	33.21	11.1	-66.6%	105.4	45.5	-56.8%
gasId	171.37	42.12	-75.4%	486.9	142.4	-70.8%
pendigits	33.97	10.69	-68.5%	109.6	43.5	-60.3%
winered	21.87	8.85	-59.5%	72.3	38.9	-46.2%
winewhite	20.36	8.65	-57.5%	66.7	37.5	-43.8%

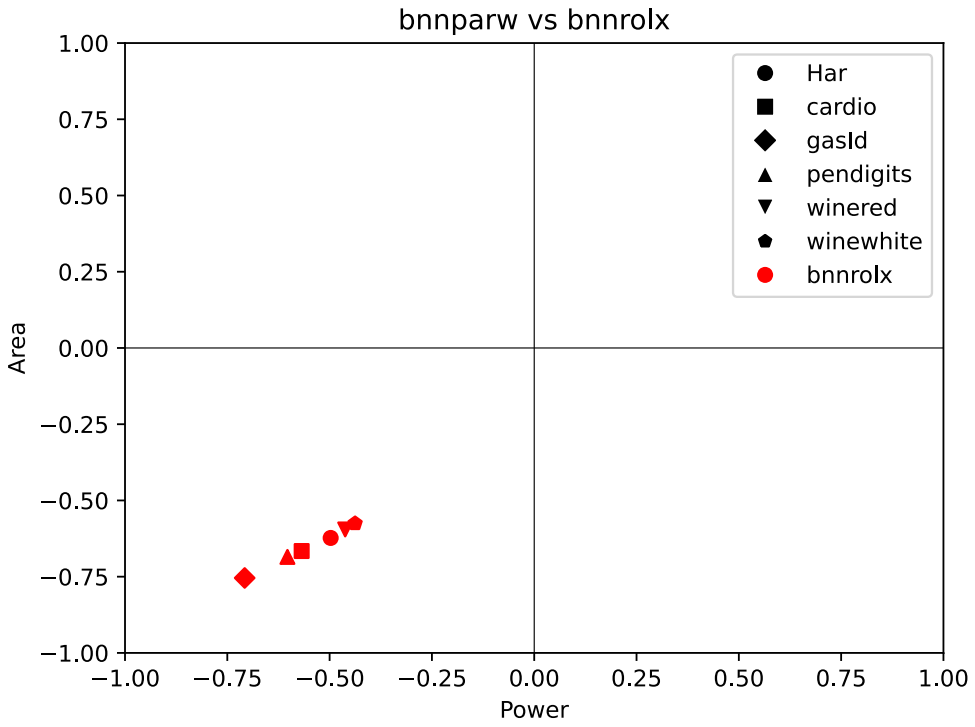

Figure 20: Σύγκριση ακολουθιακών σχεδιασμών ενιαίου δένδρου πρόσθεσης με ισοδύναμους συνδυαστικούς σχεδιασμούς..

Χρησιμοποιώντας ένα δέντρο μονών αθροιστών μας δίνει 60 - 75% μικρότερο αποτύπωμα από τα πλήρως παράλληλα κυκλώματα. Ένα κόστος σε πολυπλέκτες / αποθήκευση βάρους πρέπει να καταβληθεί εκ των προτέρων, και δεδομένου ότι τα μεγέθη των δικτύων είναι στη μικρή πλευρά η εξοικονόμηση κλίμακας από την επαναχρησιμοποίηση λογικής σε όλους τους νευρώνες δεν αποδίδει πλήρως.

7.2.2 Αποδόμηση της αρνητικής εισόδου

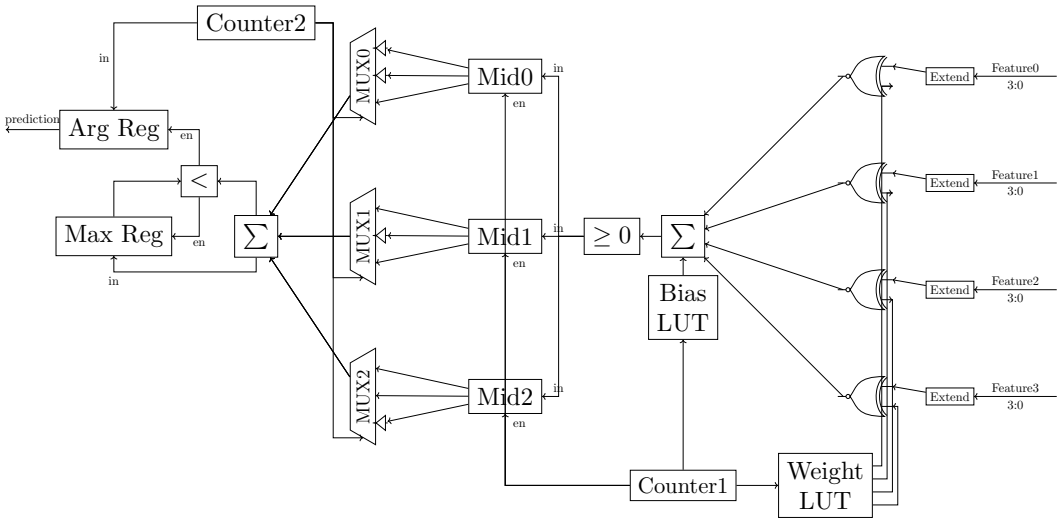

Figure 21: Οι πολυπλέκτες αντικαθίστανται από πίνακες αναζήτησης για τα βάρη και τους όρους διόρθωσης

Η άρνηση κάθε εισόδου περιλαμβάνει ένα κύκλωμα 4 bit increment-by-1 ανά χαρακτηριστικό. Αν και δεν ακούγεται πολύ ανησυχητικό, εξακολουθεί να είναι ένα έξοδο που κλιμακώνεται με τον αριθμό των εισόδων N. Η πράξη της άρνησης μπορεί να αποδομηθεί σε αντιστροφή της εισόδου και προσθήκη 1 στο αποτέλεσμα. Αν αντί να παρέχεται η αρνητική τιμή της εισόδου ως αποτέλεσμα του πολλαπλασιασμού με το κατάλληλο βάρος στον πολυπλέκτη παρέχεται η αντίστροφη τιμή όλων των bits της εισόδου, ολόκληρο το τμήμα απλοποιείται σε έναν πίνακα αναζήτησης 1 bit με δείκτη τον μετρητή κύκλων του οποίου η έξοδος γίνεται XOR με όλα τα bits του χαρακτηριστικού εισόδου. Αυτό εξοικονομεί κάποια λογική.

Για να μην υπάρχει κίνδυνος σφάλματος από αυτή την προσέγγιση, πρέπει να προστεθεί στο άθροισμα ένας όρος διόρθωσης b_i, ίσος με τον αριθμό των 1 που δεν προστέθηκαν για να αναιρεθούν σε αυτόν τον κύκλο ή με τον αριθμό των στοιχείων της γραμμής βαρών που ανήκουν στον τρέχοντα υπολογιζόμενο κρυφό νευρώνα που είναι -1.

$$b_i = \sum_{j \in W1_i} [j = -1]$$

$$h_i = \sum_{j=0}^{N-1} x_j \oplus \neg bin(W1_{i,j}) + b_i$$

7.2.3 Καταχωρητές ολίσθησης για τη χρονομέτρηση

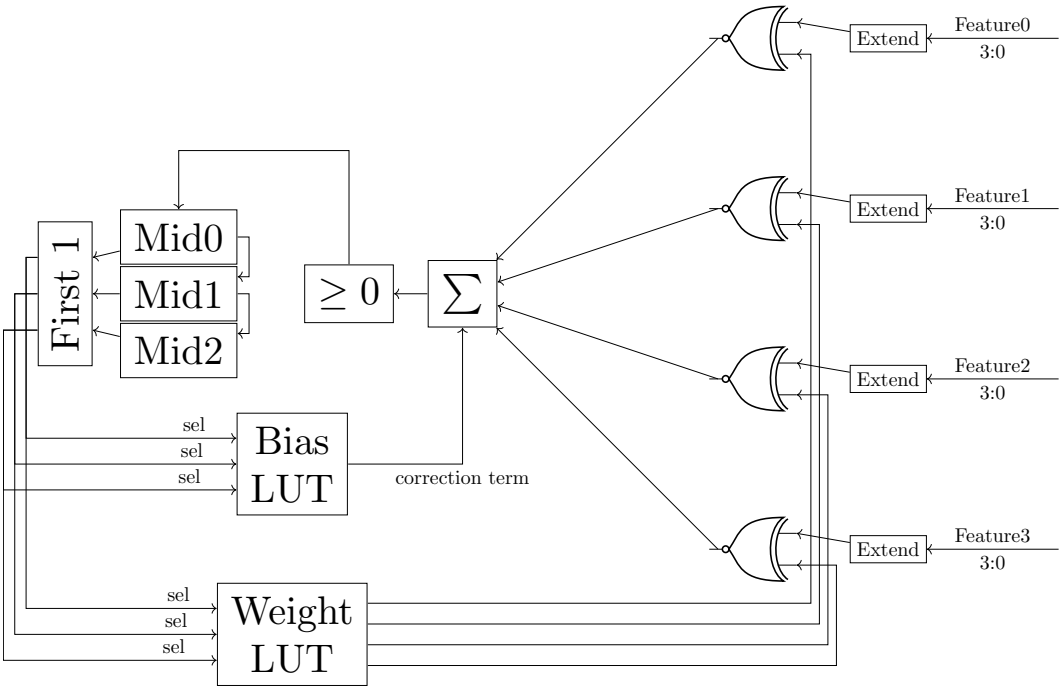

Figure 22: Αντικατάσταση του μετρητή κύκλων με καταχωρητές ολίσθησης για τις κρυφές ενεργοποιήσεις.

Αντί ένας αποκωδικοποιητής από την τρέχουσα τιμή i του μετρητή κύκλων να επιλέγει τον καταχωρητή στον οποίο θα αποθηκευτεί η δυαδική έξοδος s_i μετά την ενεργοποίηση του νευρώνα που υπολογίστηκε, είναι απλούστερο να χρησιμοποιηθεί ένας καταχωρητής ολίσθησης. Σε κάθε κύκλο οι τιμές που κρατήθηκαν προηγουμένως μετατοπίζονται κατά μία θέση προς τα δεξιά και το αποτέλεσμα της τρέχουσας αξιολόγησης αποθηκεύεται στην πιο αριστερή θέση του καταχωρητή. Αφού περάσουν N κύκλοι, η δεξιότερη θέση του καταχωρητή περιέχει το αποτέλεσμα της αξιολόγησης του πρώτου νευρώνα που έχει μετατοπιστεί κατά $N - 1$ θέσεις και όλες οι έξοδοι βρίσκονται στη σωστή τους θέση. Η σημαία που σταματά τη λειτουργία του πρώτου στρώματος τίθεται τότε και τα αποτελέσματα παγώνουν στη θέση τους για να χρησιμοποιηθούν από το επόμενο στρώμα.

Η ενσωμάτωση ενός καταχωρητή μετατόπισης, όπου οι τιμές που έχουν οριστεί κατά την αρχικοποίηση απορρίπτονται, παρέχει την ευκαιρία να καταργηθεί

εντελώς ο μετρητής κύκλων. Όταν το σήμα reset επαναφέρει τους καταχωρητές στις προκαθορισμένες τιμές τους, αναθέτουμε στο αριστερότερο bit το 1 και σε όλα τα υπόλοιπα το 0. Η θέση του πιο σημαντικού 1 στον καταχωρητή μετατοπίζεται μία φορά προς τα δεξιά σε κάθε κύκλο. Χρησιμοποιώντας ένα απλό ισοδύναμο ενός κωδικοποιητή προτεραιότητας one hot παράγεται ένα σήμα M bit μιας one hot αναπαράστασης του τρέχοντος κύκλου.

Αυτό το σήμα one-hot μπορεί να χρησιμοποιηθεί για να ανιχνεύσει πότε αξιολογείται ο τελικός νευρώνας και πρέπει να τεθεί η σημαία μετάβασης στρώματος, και μπορεί να επιλέξει τη στήλη βαρών για τη δεδομένη στιγμή που υπολογίζεται από τον πίνακα αναζήτησης χωρίς να απαιτείται αποκωδικοποιητής από τον μετρητή κύκλων. Αυτό επιτρέπει στον μετρητή να αποσυρθεί, χωρίς να απαιτούνται επιπλέον στοιχεία διατήρησης κατάστασης για τη διατήρηση της λειτουργικότητάς του.

7.2.4 Μνήμη βάρους τριών καταστάσεων

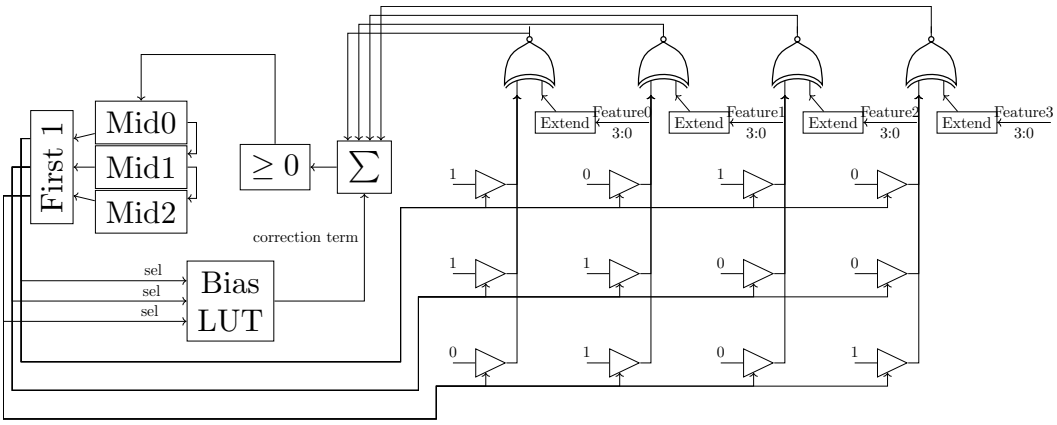

Figure 23: Υλοποίηση της μνήμης βάρους με δείκτη ενός σημείου με χρήση ανοικτού διαύλου ανά χαρακτηριστικό εισόδου

Κάθε γνώρισμα εισόδου λαμβάνει το τρέχον bit βάρους από έναν ανοικτό δίαυλο με τον οποίο συνδέεται μια tristate buffer για κάθε εγγραφή στη στήλη του χαρακτηριστικού στον πίνακα βάρους. Κάθε buffer αντιστοιχεί σε ένα μόνο στοιχείο του πίνακα βάρους $W1$. Ο tri-buffer που κρατάει την τιμή του $W1_{i,j}$ έχει την έξοδό του συνδεδεμένη στον ίδιο ανοικτό δίαυλο με τους άλλους buffers που κρατούν ένα βάρος στο $W1_{:,j}$ και ενεργοποιείται από το i-οστό bit του σήματος one-hot select από τα παραπάνω.

Ο στόχος αυτού είναι να αποφευχθούν οι εμφωλευμένες πύλες OR που χρησιμοποιούνται για την αναγωγή της επιλεγμένης τιμής της στήλης σε ένα bit στην τυπική υλοποίηση του πίνακα αναζήτησης.

7.2.5 Αποτελέσματα και ανάλυση

Table 8: Σύγκριση των τελικών ακολουθιακών σχεδίων με τα τελικά συνδυαστικά σχέδια

	bnnparw area(cm^2)	bnnrospine area(cm^2)	area change	bnnparw power(mW)	bnnrospine power(mW)	power change
Har	24.25	7.82	-67.8%	77.6	31.7	-59.1%
cardio	33.21	9.3	-72.0%	105.4	36	-65.8%
gasId	171.37	37.31	-78.2%	486.9	124.4	-74.5%
pendigits	33.97	9.08	-73.3%	109.6	35.1	-68.0%
winered	21.87	7.61	-65.2%	72.3	30.9	-57.3%
winewhite	20.36	7.49	-63.2%	66.7	30.9	-53.7%

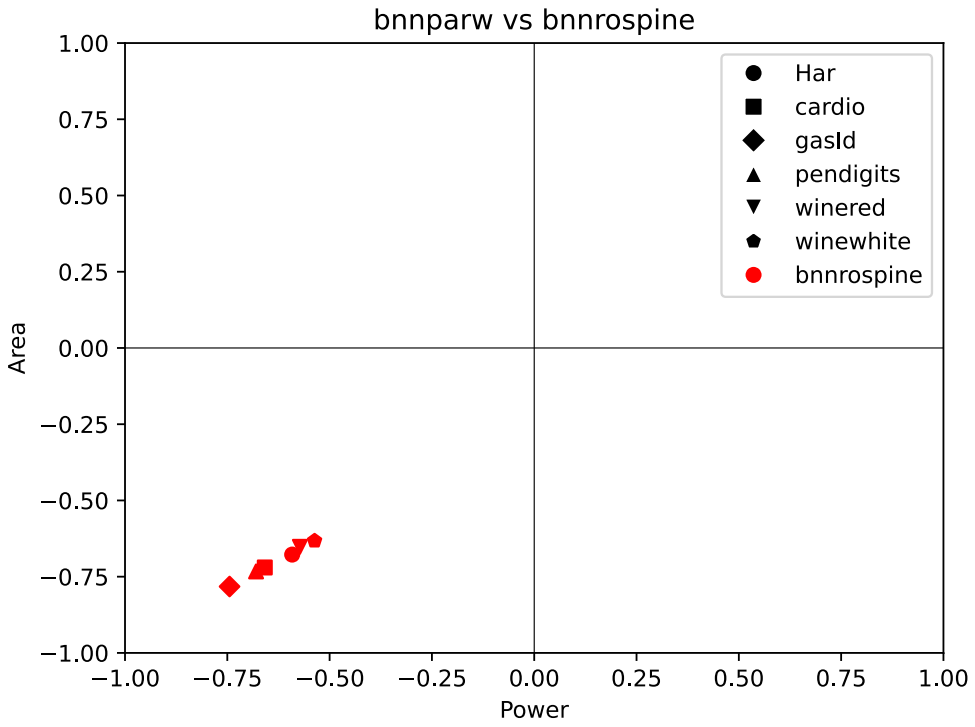

Figure 24: Σύγκριση των τελικών ακολουθιακών σχεδίων με τα τελικά συνδυαστικά σχέδια

Αυτές οι αλλαγές παρέχουν μείωση 10-20% στις απαιτήσεις σε επιφάνεια και ισχύ σε σύγκριση με την αρχική υλοποίηση του ενιαίου δέντρου αθροιστή. Με τη χρήση των ρυθμιστών τριών καταστάσεων η εξοικονόμηση ισχύος ανεβαίνει στο ~35%, αλλά με μια βαρειά ποινή έκτασης 10-30% σε σχέση με την πρώτη σχεδίαση. Νομίζω ότι η εξήγηση γι' αυτό είναι ότι απαιτούνται περισσότεροι tristate buffers από τις πύλες OR αφού δεν μπορούν να γίνουν λογικές απλοποιήσεις σε αυτούς, αλλά αφαιρείται η σημαντική κατανάλωση ισχύος μεταγωγής από τα ενδιάμεσα δίκτυα που συνδέουν τις φωλιασμένες πύλες OR. Αυτός ο συμβιβασμός επιτρέπει τη βελτιστοποίηση για όποιο από τα δύο μεγέθη της περιοχής και της ισχύος είναι το μεγαλύτερο εμπόδιο στην επιθυμητή εφαρμογή.

Το πιο σημαντικό για τους σκοπούς μας, κανένα από τα μοντέλα δεν θα μπορούσε να τροφοδοτηθεί από μια μπαταρία Molex 30mW χρησιμοποιώντας ένα συμβατικό LUT για τα βάρη. Μετά την υλοποίηση του LUT με τη χρήση tristate buffers, 5

από τα 6 μπορούν να τροφοδοτηθούν από αυτό. Αν και δεν έρχονται φθηνά από άποψη έκτασης, η εξοικονόμηση ισχύος ήταν κρίσιμη για την υπέρβαση αυτού του εμποδίου.

Συνολικά, σε σύγκριση με τα πλήρως παράλληλα σχέδια, οι απαιτήσεις μειώνονται κατά $3-5\times$. Αυτό ανοίγει τον χώρο των υλοποιήσιμων εφαρμογών. Η σχετική εξοικονόμηση θα βελτιωθεί σημαντικά για μεγαλύτερα δίκτυα, δεδομένης της παρατηρούμενης κλιμάκωσης.

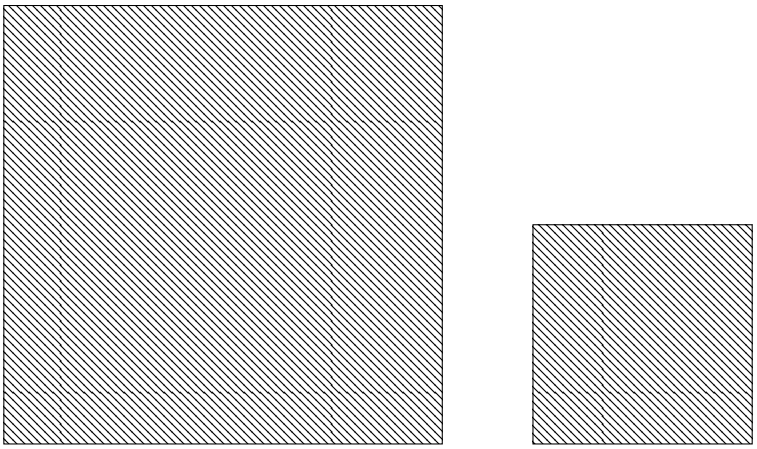

Figure 25: Σύγκριση του πραγματικού μεγέθους της εκτιμώμενης έκτασης των εκτυπωμένων σχεδίων για το μοντέλο του συνόλου δεδομένων pendigits. Παράλληλη στα αριστερά, ακολουθιακή στα δεξιά.

8 Δίκτυα τριαδικού βάρους

8.1 Συλλογισμός

Τα Τριαδικά Νευρωνικά Δίκτυα (**TNNs**) χρησιμοποιούν βάρη και ενεργοποιήσεις στο εύρος {-1, 0, 1} αντί του {-1, 1} των BNNs. Αυτό επιτρέπει πολύ μεγαλύτερη αναπαραστατική ικανότητα και έτσι επιτυγχάνονται υψηλότερες ακρίβειες. όταν υπολογίζονται με CPUs ή GPUs η συμπερίληψη του 0 καθιστά τις πράξεις σε επίπεδο bit που καθιστούν τα BNNs τόσο φιλικά προς τους υπολογισμούς μη εφαρμόσιμες, αφού οι ενεργοποιήσεις και τα βάρη καταλαμβάνουν τώρα δύο bits και οι πράξεις MAC τους δεν μπορούν να αναχθούν σε XNORs και popcounts. Με το υλικό που προορίζεται για την εκτέλεση των TNNs και υλοποιείται σε FPGAs ή ASICs μπορούν να γίνουν περισσότερες βελτιστοποιήσεις, αλλά η αποδοτικότητα εξακολουθεί να είναι υποδεέστερη σε σύγκριση με τα δυαδικά δίκτυα.

Στην περίπτωσή μας τα σχέδια είναι πλήρως προσαρμοσμένα σε ένα μόνο μοντέλο/σύνολο βαρών. Αυτό επιτρέπει την εκμετάλλευση του ισομορφισμού μεταξύ των τριμερών δικτύων βαρών και των αραιών δυαδικών δικτύων βαρών, δεδομένου ότι οι συνδέσεις που αφαιρούνται από το αραιό δίκτυο μπορούν να παραλειφθούν από τη σχεδίαση εκ των προτέρων, κάτι που δεν είναι εφικτό στην περίπτωση που πρέπει να υποστηρίζονται όλα τα μοντέλα μιας συγκεκριμένης αρχιτεκτονικής.

Για τα πλήρως συνδυαστικά σχέδια αυτό μεταφράζεται σε λιγότερες αριθμητικές πράξεις για την κατασκευή στοιχείων. Η ακρίβεια του μοντέλου και οι απαιτήσεις του κυκλώματος που προκύπτουν σε επιφάνεια/ισχύ βελτιώνονται και οι δύο με αυτόν τον τρόπο με τη μετάβαση σε τριαδικά βάρη.

Μόνο τα βάρη αλλά όχι οι ενεργοποιήσεις θα τριαδοποιηθούν σε αυτή την εφαρμογή. Η πρόσθετη ικανότητα του μοντέλου που επιτυγχάνεται με τη χρήση τριαδικών ενεργοποιήσεων για το κρυφό στρώμα πέραν των βαρών δεν ήταν αρκετά σημαντική ώστε να δικαιολογήσει το εικαζόμενο κόστος της υλοποίησης της αριθμητικής 2-bit στο επόμενο στρώμα.

8.2 Πλήρως συνδυαστική υλοποίηση

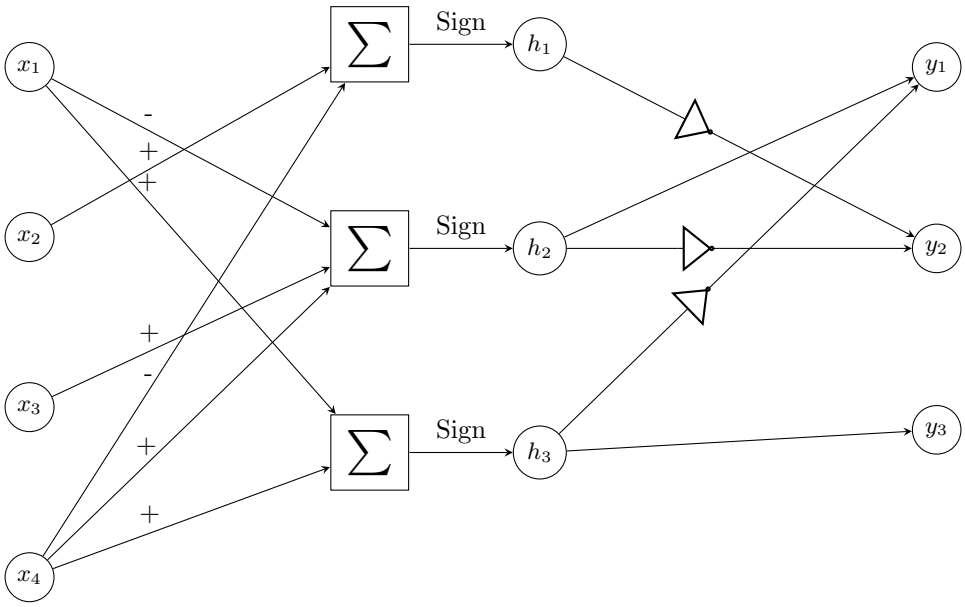

Figure 26: Sparce συνδυαστική υλοποίηση ενιαίου προσημασμένου αθροίσματος.

Αφού εκπαιδεύσουμε τα ίδια σύνολα δεδομένων με τις ίδιες μετρήσεις παραμέτρων χρησιμοποιώντας τριαδικά αντί για δυαδικά βάρη οι ισοδύναμοι πίνακες βαρών $W1 \in \{-1, 0, 1\}^{M,N}$ και $W2 \in \{-1, 0, 1\}^{C,M}$ χρησιμοποιούνται για να ορίσουμε τις επιθυμητές πράξεις παρόμοια με τους αντίστοιχους δυαδικούς:

$$h_i = \sum_{j=0}^{N-1} \begin{cases} +x_j, & \text{if } W1_{i,j} > 0 \\ -x_j, & \text{if } W1_{i,j} < 0 \end{cases}$$

$$y_i = \sum_{j=0}^{M-1} \begin{cases} s_j, & \text{if } W2_{i,j} > 0 \\ \neg s_j, & \text{if } W2_{i,j} < 0 \end{cases}$$

Στο πρώτο στρώμα για τις συνδέσεις που έχουν βάρη 0 ούτε η πρόσθεση ούτε η αφαίρεση πραγματοποιείται για το χαρακτηριστικό στην καθορισμένη έκφραση του νευρώνα.

Απλώς αγνοείται, αφού συνεισφέρει πάντα 0 στο άθροισμα.Ορισμένοι νευρώνες αποδεικνύεται ότι όλα τα μη μηδενικά βάρη τους έχουν το ίδιο πρόσημο, είτε όλα είναι 1 είτε όλα -1. Σε μια τέτοια περίπτωση, εφόσον τα χαρακτηριστικά εισόδου

είναι όλα θετικά, το πρόσημο του αποτελέσματος της πολλαπλής συσσώρευσης για αυτόν τον νευρώνα θα είναι πάντα το ίδιο. Έτσι, οι έξοδοι αυτών των νευρώνων είναι κωδικοποιημένες σε σταθερές για να αποφευχθεί η περιττή επιβάρυνση.Στο δεύτερο επίπεδο, επίσης, οι συνδέσεις με βάρη ίσα με μηδέν δεν συμπεριλαμβάνουν ούτε το κρυφό χαρακτηριστικό με το οποίο συνδέονται ούτε το αντίστροφό του στο popcount του σχετικού νευρώνα εξόδου. Η απαιτούμενη λογική μειώνεται έτσι και στα δύο στρώματα με κάθε σύνδεση που αποκόπτεται.

Υπενθυμίζεται ο γραμμικός μετασχηματισμός που χρησιμοποιήθηκε για να πάμε από τα γινόμενα ενεργοποίησης βαρών των αθροισμάτων {-1,1} σε popcounts των XNORs που τώρα κωδικοποιούν τις τιμές που θα ήταν -1 ως 0. Για να επιτευχθεί $-1 \rightarrow 0$ και $1 \rightarrow 1$ $f(x) = \frac{(x+1)}{2}$ είναι αυτός ο γραμμικός μετασχηματισμός. Όταν εφαρμόζεται σε ένα array $v \in \{-1, 1\}^M$ δυαδικών τιμών το άθροισμα είναι:

$$\sum_{i=0}^{M-1} f(v_i) = \sum_{i=0}^{M-1} (v_i + 1)/2 = \frac{1}{2} \sum_{i=0}^{M-1} v_i + 1 = \frac{1}{2} \sum_{i=0}^{M-1} v_i + \frac{M}{2}$$

Στο τελευταίο στρώμα, που σε αυτή την περίπτωση είναι το δεύτερο, μέχρι στιγμής ο αριθμός των εισόδων όλων των νευρώνων εξόδου ήταν ο ίδιος, επομένως ο παράγοντας $\frac{M}{2}$ θα μπορούσε να παραλειφθεί κατά τη σύγκριση των εξόδων μεταξύ των νευρώνων και τα αποτελέσματα XNOR/popcount μπορούν να χρησιμοποιηθούν απευθείας για τους υπολογισμούς argmax.

Τώρα, αφού οι νευρώνες εξόδου αντιμετωπίζονται ως αραιές δυαδικές μονάδες, δεν έχουν πλέον τον ίδιο αριθμό εισόδων και επομένως ο σταθερός όρος $\frac{M}{2}$ δεν είναι πλέον ο ίδιος για ολόκληρο το επίπεδο. Το ζήτημα μπορεί επίσης να περιγραφεί ως στοιχεία με τιμή 0 που περιλαμβάνονται στο διάνυσμα v και συνεισφέρουν $f(0) = \frac{1}{2}$ το καθένα στο άθροισμα μετά τη γραμμική μετατροπή, κάτι που δεν αντικατοπτρίζεται αν τα αγνοήσουμε εντελώς όπως γίνεται εδώ.

Για να διορθωθεί αυτό το ζήτημα, ένας όρος διόρθωσης ίσος με $\frac{z_i}{2}$ θα πρέπει να προστεθεί στο αποτέλεσμα που δίνει ο υπολογισμός XNOR/popcount, όπου z_i δηλώνει τον αριθμό των στοιχείων στη γραμμή βάρους του i-οστού νευρώνα εξόδου $W2_i$. Αυτός ο όρος μπορεί να είναι μεγάλος σε σχέση με την τιμή του popcount όταν ένας νευρώνας είναι αρκετά αραιός, οπότε αντί αυτού μπορούμε να προσθέσουμε $\frac{z_i - \min_{j=0}^{M-1} z_j}{2}$ ως μικρότερο διορθωτικό όρο, έτσι ώστε ο νευρώνας με τις λιγότερες αποκομμένες συνδέσεις να μην προσθέτει τίποτα στο popcount του και οι υπόλοιποι να παίρνουν το δικό τους με βάση το πόσα επιπλέον μηδενισμένα βάρη περιέχουν σε σχέση με αυτόν.

8.3 Αποτελέσματα και ανάλυση

Table 9: Απόδοση της συνδυαστικής υλοποίησης με τριαδικά βάρη σε σύγκριση με τη συνδυαστική υλοποίηση με δυαδικά βάρη

	bnnparw area(cm²)	tnnparsign area(cm²)	area change	bnnparw power(mW)	tnnparsign power(mW)	power change
Har	24.25	13.4	-44.7%	77.6	42.7	-45.0%
cardio	33.21	19.21	-42.2%	105.4	62.4	-40.8%
gasId	171.37	101.65	-40.7%	486.9	297.1	-39.0%
pendigits	33.97	29.43	-13.4%	109.6	95.8	-12.6%
winered	21.87	11.78	-46.1%	72.3	40	-44.7%
winewhite	20.36	9.53	-53.2%	66.7	32.8	-50.8%

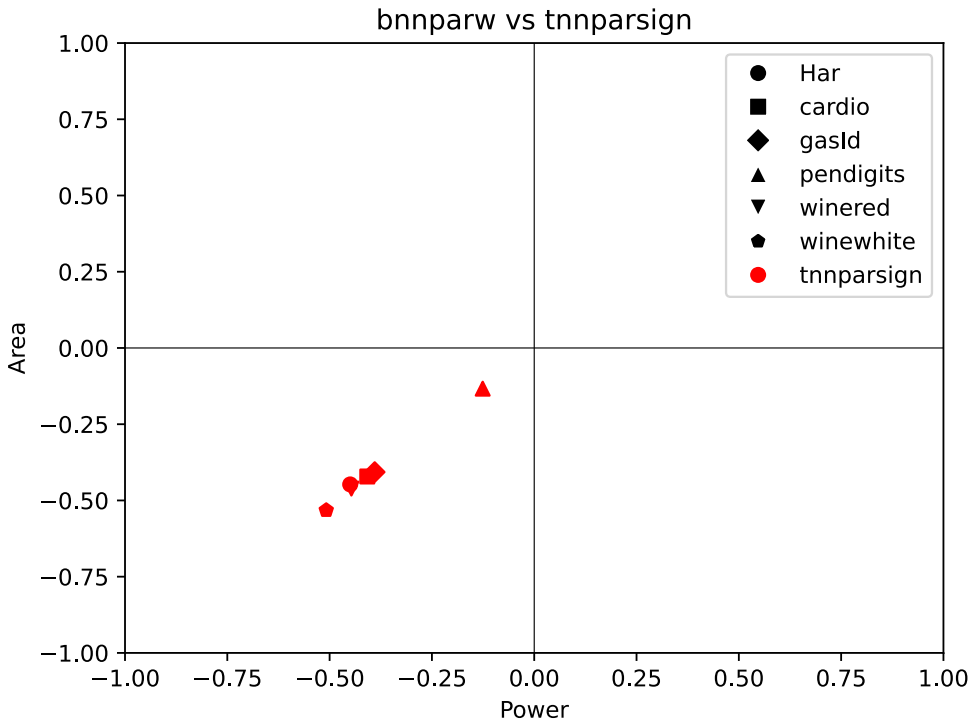

Figure 27: Απόδοση της συνδυαστικής υλοποίησης με τριαδικά βάρη σε σύγκριση με τη συνδυαστική υλοποίηση με δυαδικά βάρη

Χάρη στην αφαίρεση των όρων από τις αριθμητικές πράξεις που καθορίζουν το σχεδιασμό των στρωμάτων οι απαιτήσεις σε επιφάνεια και ισχύ ήταν σχεδόν στο μισό σε σύγκριση με τους πλήρως συνδυαστικούς σχεδιασμούς των δυαδικών μοντέλων που εκπαιδεύτηκαν στα ίδια σύνολα δεδομένων με τα τριμερή. Η ακρίβεια βελτιώθηκε επίσης σε όλους τους τομείς.

Δυστυχώς, η εφαρμογή της μείωσης του bitwidth ή της προ της σύνθεσης αριθμητικής βελτιστοποίησης υπολειπόταν των αρχικών κυκλωμάτων TNN, ακόμη και για σύνολα δεδομένων των οποίων οι υλοποιήσεις δυαδικών δικτύων βελτιώθηκαν με αυτές τις μεθόδους. Οι απόπειρες ακολουθιακών σχεδιασμών για δίκτυα τριαδικών βαρών δεν αποδίδουν σε ικανοποιητικό επίπεδο, οπότε δεν θα επεκταθούμε σε αυτές.

[1] IDTechEx, *Flexible & printed electronics 2023-2033: Forecasts, technologies, markets*. 2023.

[2] V. Sze, Y. Chen, T. Yang, and J. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, pp. 2295–2329, 2017.

[3] D. Weller, M. Hefenbrock, M. Tahoori, J. Aghassi-Hagmann, and M. Beigl, "Programmable neuromorphic circuit based on printed electrolyte-gated transistors," in *2020 25th asia and south pacific design automation conference (asp-dac)*, 2020, pp. 446–451.

[4] M. Douthwaite, F. García-Redondo, P. Georgiou, and S. Das, "A time-domain current-mode mac engine for analogue neural networks in flexible electronics," in *2019 ieee biomedical circuits and systems conference (biocas)*, IEEE, 2019, pp. 1–4.

[5] H. Ling, D. Koutsouras, S. Kazemzadeh, Y. van de Burgt, F. Yan, and P. Gkoupidenis, "Electrolyte-gated transistors for synaptic electronics, neuromorphic computing, and adaptable biointerfacing," *Applied Physics Reviews*, vol. 7, no. 1, p. 011307, 2020.

[6] E. Ozer *et al.*, "Bespoke machine learning processor development framework on flexible substrates," in *2019 ieee international conference on flexible and printable sensors and systems (fleps)*, IEEE, 2019, pp. 1–3.

[7] N. Bleier, M. Mubarik, F. Rasheed, J. Aghassi-Hagmann, M. Tahoori, and R. Kumar, "Printed microprocessors," in *2020 acm/ieee 47th annual international symposium on computer architecture (isca)*, IEEE, 2020, pp. 213–226.

[8] D. Weller *et al.*, "Printed stochastic computing neural networks," in *Design, automation test in europe conference exhibition (date)*, 2021, pp. 914–919.

[9] M. Mubarik *et al.*, "Printed machine learning classifiers," in *Annu. Int. Symp. Microarchitecture (micro)*, 2020, pp. 73–87.

[10] G. Armeniakos, G. Zervakis, D. Soudris, M. Tahoori, and J. Henkel, "Cross-layer approximation for printed machine learning circuits," in *Design, automation test in europe conference exhibition (date)*, 2022. Available: https://arxiv.org/abs/2203.05915

[11] A. Kokkinis, G. Zervakis, K. Siozios, M. B. Tahoori, and J. Henkel, "Hardware-aware automated neural minimization for printed multilayer perceptrons," in *2023 design, automation & test in europe conference & exhibition (date)*, IEEE, 2023, pp. 1–2.

[12] G. Armeniakos, G. Zervakis, D. Soudris, M. B. Tahoori, and J. Henkel, "Model-to-circuit cross-approximation for printed machine learning classifiers," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.

[13] G. Armeniakos, G. Zervakis, D. Soudris, M. B. Tahoori, and J. Henkel, "Co-design of approximate multilayer perceptron for ultra-resource constrained printed circuits," *IEEE Transactions on Computers*, 2023.

[14] K. Balaskas, G. Zervakis, K. Siozios, M. B. Tahoori, and J. Henkel, "Approximate decision trees for machine learning classification on tiny printed circuits," in *2022 23rd international symposium on quality electronic design (isqed)*, IEEE, 2022, pp. 1–6.

[15] K. Iordanou *et al.*, "Tiny classifier circuits: Evolving accelerators for tabular data." 2023. Available: http://arxiv.org/abs/2303.00031

[16] G. Cadilha Marques, D. Weller, A. T. Erozan, X. Feng, M. Tahoori, and J. Aghassi-Hagmann, "Progress report on 'from printed electrolyte-gated metal-oxide devices to circuits'," *Advanced Materials*, vol. 31, no. 26, p. 1806483, 2019.

[17] L. Shao, T.-C. Huang, T. Lei, Z. Bao, R. Beausoleil, and K.-T. Cheng, "Compact modeling of carbon nanotube thin film transistors for flexible circuit design," in *2018 design, automation & test in europe conference & exhibition (date)*, IEEE, 2018, pp. 491–496.

[18] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," *arXiv preprint arXiv:1602.02830*, 2016.

[19] M. Kim and P. Smaragdis, "Bitwise neural networks," *arXiv preprint arXiv:1601.06071*, 2016.

[20] D. Dua and C. Graff, "UCI machine learning repository." http://archive.ics.uci.edu/ml, 2017.

[21] H. A. Guvenir, B. Acar, G. Demiroz, and A. Cekin, "A supervised machine learning algorithm for arrhythmia analysis," in *Computers in cardiology 1997*, IEEE, 1997, pp. 433–436.

[22] D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de Sa, and L. Pereira-Leite, "Sisporto 2.0: A program for automated analysis of cardiotocograms," *Journal of Maternal-Fetal Medicine*, vol. 9, no. 5, pp. 311–318, 2000.

[23] F. Alimoglu and E. Alpaydin, "Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition," in *Proceedings of the fifth turkish artificial intelligence and artificial neural networks symposium (tainn 96)*, Citeseer, 1996.

[24] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Esann*, 2013.

[25] S. Feng *et al.*, "Review on smart gas sensing technology," *Sensors*, vol. 19, no. 17, p. 3760, 2019.

[26] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

[27] E. Bihar, T. Roberts, M. Saadaoui, T. Hervé, J. B. De Graaf, and G. G. Malliaras, "Inkjet-printed pedot:PSS electrodes on paper for electrocardiography," *Advanced Healthcare Materials*, vol. 6, no. 6, 2017.

[28] J. Dai *et al.*, "Printed gas sensors," *Chemical Society Reviews*, vol. 49, no. 6, pp. 1756–1789, 2020.

[29] S. Tuukkanen and S. Rajala, "A survey of printable piezoelectric sensors," in *2015 ieee sensors*, IEEE, 2015, pp. 1–4.

[30] M. Jose *et al.*, "Printed pH sensors for textile-based wearables: A conceptual and experimental study on materials, deposition technology, and sensing principles," *Advanced Engineering Materials*, vol. 24, no. 5, 2022.

[31] M. Jelbuldina, H. Younes, I. Saadat, L. Tizani, S. Sofela, and A. Al Ghaferi, "Fabrication and design of cnts inkjet-printed based micro fet sensor for sodium chloride scale detection in oil field," *Sensors and Actuators A: Physical*, vol. 263, pp. 349–356, 2017.

[32] S. Ma and P. Ampadu, "Optimal sat-based minimum adder synthesis of linear transformations," in *2019 ieee 62nd international midwest symposium on circuits and systems (mwscas)*, IEEE, 2019, pp. 335–338.

[33] C. Paar, "Optimized arithmetic for reed-solomon encoders," in *Proceedings of ieee international symposium on information theory*, IEEE, 1997, p. 250.