

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – ВАРНА

инж. Николай Тинков Дуков

Заглавие:

**ИЗСЛЕДВАНЕ НА ВЪЗМОЖНОСТИ ЗА СЪЗДАВАНЕ НА
ИМИТАЦИОННИ БАЗОВИ НЕВРОННИ СТРУКТУРИ И
ТЕХНИТЕ РЕАЛИЗАЦИИ**

АВТОРЕФЕРАТ

**на дисертация за получаване на образователна и
научна степен „доктор“**

**по докторска програма: 5.2.4 „Електронизация“
към професионално направление: 5.2 „Електротехника, електроника
и автоматика“**

Научни ръководители:

доц. д-р инж. Димитър Михайлов Ковачев
доц. д-р инж. Тодор Димитров Ганчев

Рецензенти:

1.
2.

**Варна
2018 г.**

Дисертационният труд е обсъден на 09.11.2018 г. в катедра „Електронна техника и микроелектроника“ и е насочен за защита.

Докторантът работи в катедра „Компютърни науки и технологии“.

Автор: инж. Николай Тинков Дуков

Заглавие: Изследване на възможности за създаване на имитационни базови невронни структури и техните реализации

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – ВАРНА

инж. Николай Тинков Дуков

Заглавие:

**ИЗСЛЕДВАНЕ НА ВЪЗМОЖНОСТИ ЗА СЪЗДАВАНЕ НА
ИМИТАЦИОННИ БАЗОВИ НЕВРОННИ СТРУКТУРИ И
ТЕХНИТЕ РЕАЛИЗАЦИИ**

АВТОРЕФЕРАТ

**на дисертация за получаване на образователна и
научна степен „доктор“**

**Варна
2018 г.**

Дисертационният труд съдържа 146 страници, включително 43 фигури и 30 таблици, оформени в 5 глави, общи изводи и списък на използваната литература от 137 заглавия, от които 3 на кирилица и 134 на латиница.

Защитата на дисертационния труд ще се състои на г. от ч. в на открито заседание на жури, сформирано със заповед на Ректора №/..... г.

Материалите по защитата (дисертация, рецензии и становища) са на разположение на интересуващите се в Докторантски център, стая 318 НУК

ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

Голяма част от съвременните изкуствени невронни мрежи са реализирани като програмно приложение за изчислителни машини с общо предназначение. Съществуват и множество приложения в универсални и специализирани преносими устройства, като медицински монитори, свръх-леки автоматични регистратори на животински акустични емисии и др., при които се изисква автоматизирано регистриране на събития. Бърз и лесен начин за реализиране на една такава функционалност е зареждане на програмно приложение на готова, достъпна апаратна платформа с популярна операционна система, като например интелигентен телефон [1], [2], [3], [4], [5], [6]. Сред недостатъците на този подход са ниската надеждност и ниската ефективност.

Проблем

Използването на реконфигурируеми апаратни устройства предоставя добра комбинация между възможностите на специализирано апаратно устройство и програмно приложение. Електрическите програмируеми матрични кристали (FPGA) са вид апаратни логически устройства, които имат свойството да бъдат препрограмирани като изчислителна платформа с общо предназначение, но въпреки това да запазват високото си бързодействие, сравнимо с това на интегрални схеми със специално приложение. Предимствата при реализацията на алгоритми за машинно обучение, като намалена консумация и подобрена стабилност на работа стават явни, особено при сравнението между различни платформи, като CPU, GPU и FPGA при осигурена еквивалентна изчислителна работа [7], [8].

Цел и задачи на изследването

Цел:

Създаване на модели на невронни структури за класификация и методи за обучението им, с насоченост към ефективна апаратна реализация с FPGA технологии.

Задачи:

1. Изследване и анализ на възможностите за реализиране на невронни структури за класификация с FPGA технологии.
2. Създаване на удобни за апаратна реализация модели на невронни структури с насоченост към конкретни типови задачи (бинарни изображения, многоканални ЕЕГ сигнали, речеви сигнали).
3. Оптимизиране на създадените модели на невронни структури за нуждите на конкретни приложения (разпознаване на сърдечни аномалии, откриване на болест на паркинсон, разпознаване на негативни емоционални състояния).

4. Изследване на методи за обучение на класификатори, изградени с невронни структури в апаратна реализация.

Обект и предмет на изследване

Обект на изследване са невронни структури с особености и характеристики, които ги правят подходящи за апаратна реализация, както и съответните методи за тяхното обучение.

Предмет на научно изследване са методите за реализиране на невронни структури с FPGA и адаптацията към спецификата на конкретни приложения. Проучени са възможностите за адаптиране, както и за създаването на нови архитектури удобни за апаратна реализация. Изследвана е пригодността на различни методи за обучение на реализираните невронни структури.

Методи на изследване

Методите, чрез които са постигнати поставените задачи в дисертацията, включват:

- Статистически и аналитични методи за апроксимация на функции.
- Моделиране и симулация на невронни структури реализирани в програмна и апаратна среда.
- Експериментални изследвания за верифициране на предложените модели.
- Статистически методи за обективна качествена и количествена оценка на получените резултати от експерименталните изследвания.

Място на изследване

Изследванията, представени в дисертацията, са проведени на следните места:

- Технически Университет – Варна, България – „Лаборатория по приложна обработка на сигнали“;
- Патренски Университет, Патра, Гърция – “Лаборатория по изчислителен интелект” (CILab), в рамките на студентка програма за мобилност „Еразъм“.

Научна новост на изследването

Изследването представено в дисертационния труд предоставя нови идеи за реализация на вероятностната невронна мрежа (ВНМ) и локално рекурентната вероятностна невронна мрежа (ЛРВНМ), като и методи за адаптация на определени невронни структури с цел подобряване на ефективността на тяхната реализация с апаратни средства (FPGA). Създадена е нова разновидност на ЛРВНМ, която в локално рекурентния слой използва неврони с активираща функция от семейството на ReLU (Rectified Linear Unit). Създаденият нов модел е удобен за апаратна реализация и е с подобрена точността на класификация спрямо традиционният модел. Създаден е паралелен апаратен модел на ЛРВНМ, с висока разрядност на предавателната функция в рекурентния слой, който постига точност на класификация

сходна с тази на програмния модел, но при многократно подобрене на бързодействието. Предложени са голям брой оптимизации на модела на ВНМ, позволяващи редуциране на използваните FPGA ресурси при класификация на бинарни изображения.

Практическа ценност на изследването

Създаденият нов модел на неврон за рекурентния слой на ЛРВНМ предоставя възможности за по-нататъшното развитие на локално-рекурентната архитектура. Разработените приложно-оптимизирани апаратни модели на ВНМ и ЛРВНМ могат да подпомогнат разширяване обхвата на приложенията им. Представената експериментална оценка на различни методи за обучение на ВНМ е полезна при избор на обучаващ метод за конкретно приложение. Предложените методи, апаратни модели, и неврони създават предпоставки за използването на FPGA реализации на ВНМ и ЛРВНМ в интелигентни, малки, евтини, преносими устройства с ниска консумация и дълъг живот.

Апробация на изследването

Основните теоретични и приложни резултати от дисертационния труд са представени в общо 10 публикации, от които:

- 2 доклада на конференция „UNITECH“, 2014 г., 2015 г., Габрово, България;
- 2 доклад в годишник на ТУ-Варна, 2014 г., 2015 г.;
- 1 доклад на научен форум „Иновации и Бизнес 2015“, 2015 г., Варна, България;
- 1 доклад на конференция „IEEE XXV International Scientific Conference Electronics – ET2016“, 2016 г., Созопол, България;
- 1 доклад на симпозиум „Biomedical data acquisition and applications workshop“, 2017 г., Варна, България;
- 1 доклад на конференция „Intelligent information technologies for industry – ИТИ2017“, 2017 г., Варна, България;
- 1 доклад на конференция „V-та научна конференция с международно участие "Компютърни науки и технологии"“, 2018 г., Варна, България;
- 1 доклад в списание „Expert Systems With Applications“, Elsevier (в процес на преразглеждане (under revision)).

Конференциите „IEEE ET2016“, „ИТИ2017“ и списанието „Expert Systems With Applications“ са индексирани в международната научна база данни „SCOPUS“.

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Глава 1: Теоретични основи и анализ на резултатите от досегашни научни изследвания по проблема

В Глава 1 са представени спецификите и предизвикателствата при реализацията на изкуствени невронни мрежи, анализирани и изследвани са възможностите за реализирането им с апаратни средства, и е представена избраната технология за реализация.

1.1. Изкуствени невронни мрежи

По своето устройство изкуствените невронни мрежи се състоят от паралелни, свързани изчислителни единици наречени неврони. Всеки изкуствен неврон получава сигнал отвън или от друг неврон, а при възбуждане изпраща сигнал до всички други свързани неврони. Възбуждането на изкуствения неврон се контролира чрез функция, известна като активираща функция [13].

1.1.1. Активиращи функции

Съществуват различни типове активиращи функции, като едни от най-често използваните са [13]: линейна функция, прагова функция, рампова функция, сигмоидална функция, хиперболична тангенс функция, Гаусова функция.

1.1.2. Видове невронни мрежи

В съвременните невронни мрежи, невроните са организирани в слоеве. Най-често това са входен слой, приемащ информация за средата; n на брой скрити слоеве, реализиращи функционалността на невронната мрежа и изходен слой, който формира отговора на мрежата. Невроните в един слой могат да бъдат изцяло или частично свързани с невроните в друг слой, като е възможно да съществуват и връзки до неврони от предишни слоеве.

Спрямо архитектурата могат да се открият три основни категории изкуствени невронни мрежи [13], [14]: еднопосочна еднослойна невронна мрежа, еднопосочна многослойна невронна мрежа, рекурентни невронни мрежи.

Обучението на изкуствените невронни мрежи се състои в настройката на тегловните коефициенти и отместването, до достигане на определен критерий. Основните типове обучение се разделят на три вида [13], [14]: контролирано обучение, подсилено обучение, неконтролирано обучение.

1.2. Електрически програмируеми матрични кристали

Електрическите програмируеми матрични кристали (FPGA) постигат скорости на изчисление сравними с тези на интегралните схеми със специално приложение (ASIC). За разлика от ASIC, FPGA притежават гъвкавостта на програма.

За постигане на реконфигурируемост, всеки чип с FPGA трябва да съдържа три основни компонента – логически блокове, взаимосвързани елементи и входно-изходни клетки [34], [31]. Днешните FPGA съдържат също блокове с памет, аритметични блокове, блокове за управление на тактовия сигнал, комуникационни блокове и други, като всички те са монолитно вградени [32], [30].

1.3. Апаратни изкуствени невронни мрежи

Реализацията на невронна мрежа в апаратен вид предлага значителни преимущества [41], [42], [43], [44], като основно преимущество е изчислителната скорост. Апаратна реализация специализирана за конкретно приложение може лесно да надмине няколкократно бързодействието на съвременен персонален компютър или цифров сигнален процесор [45]. Това твърдение важи с още по-голяма сила при естествен паралелизъм на реализирания дизайн. Друго преимущество, което не може да се пренебрегне е цената. В преносими потребителски приложения изискващи обработка на информация в реално време, по-малкия брой компоненти и занижената консумация на енергия оказват съществен принос за намаляване на общата себестойност [41].

В настоящия дисертационен труд изследванията са насочени към реализация на невронни структури в апаратна среда реализирана с FPGA. FPGA показват балансирани характеристики спрямо други популярни платформи за реализация на невронни мрежи, като аналогова ASIC, цифрова ASIC, процесор базирани, паралелни компютри. Основното предизвикателство при реализация на изкуствените невронни мрежи с FPGA е големият обем аритметични операции, които изискват значителни апаратни ресурси.

1.4. Обобщения и изводи

От направения анализ са изведени преимуществата на FPGA при апаратна реализация на изкуствени невронни мрежи: гъвкавостта на програмната реализация, добри изчислителни възможности, ефективност и бързодействие сравними с ASIC.

При апаратната реализация на изкуствени невронни мрежи основен проблем се явява ограничения ресурс на FPGA. Извършеният литературен обзор показва, че гъвкавостта на тази технология е и в известна степен пречка, тъй като много от реализираните архитектури биват реализирани директно, без или с минимален опит за опростяване на описващите ги аритметични операции. Именно аритметичните операции съставляват основна част от консумираните ресурси на FPGA чипа. Също така, анализът показва, че често малко внимание се обръща на начина на представяне на данните. Един по-индивидуален подход към дадено приложение би довел до по-прецизно обосновани компромиси между точност, бързодействие и заета площ в чипа.

1.5. Използвани инструменти и ресурси

1.5.1. Развойна среда

- За реализиране на невронните структури в програмно изпълнение е използван програмният продукт MATLAB като развойна среда. Отново в MATLAB са реализирани оптимизационните алгоритми за обучението на разгледаните невронни структури [98].
- За описание на невронните структури като цифрова логика е използван програмният продукт Quartus II и език за описание на хардуер Verilog HDL. За верификацията на описаните дизайни е използван програмният продукт ModelSim [37].
- За реализация на невронните структури във FPGA е използвана апаратната развойна среда DE2-115, с FPGA чип Altera Cyclone IV [99], [37].

1.5.2. Базы данни

- Откриване на сърдечни заболявания от пациентски биомедицински изображения (SPECT Heart Data Set) [100], [101]. Базата данни се състои от 267 изображения от еднофотонна емисионна компютърна томография. От изображенията са извлечени обобщаващи признаци и всяко от изображенията е класифицирано като нормално или анормално. Изображенията са допълнително обработени, при което за всяко изображение са получени 22 бинарни признака.
- Откриване на болестта на Паркинсон от реч (Parkinson Speech Data Set) [102], [103]. Базата данни се състои от записана реч от здрави и болни от Паркинсон участници, като резултат данните са разделени съответно в два класа – положителни за Паркинсон и отрицателни за Паркинсон. Записите са оценени от експерт по единичната скала за оценка на болестта на Паркинсон (UPDRS). Извлечени са 26 дескриптора.
- Откриване на емоционални състояния от ЕЕГ сигнали (Database for Emotion Analysis using Physiological signals - DEAP) [104]. Базата данни се състои от записи на тридесет и двама участника гледащи музикални клипове, като за всеки участник са снети по четиридесет записа. Записаната информация е от 40 канала, в т.ч. 32 канала електроенцефалографски сигнали. DEAP съдържа още електромиографски, електроокулографски и други сигнали снети от различни части на главата и тялото.

Глава 2: Вероятностна невронна мрежа за бинарни изображения в електрически програмируеми матрични кристали

В Глава 2 от дисертационния труд е представен адаптиран паралелен апаратен модел на вероятностната невронна мрежа (ВНМ) насочен към класификацията на бинарни изображения и са предложени стъпки целящи оптимизиране на дизайна от към необходими ресурси.

2.1. Архитектура на вероятностната невронна мрежа

ВНМ се състои от четири слоя. Първият слой е входен слой, той приема входните вектори за класификация. Във втория и третия слой, наречени слой с образци и сумиращ слой се формират вероятностите за всеки от класовете [105].

Различни функции за изчисляване на разпределението на вероятността са изследвани в Глава 3. Тук се разглежда общия случай на изчисление на тази функция, а именно чрез сумата от сферични Гаусови функции [105], [106]:

$$p_i(\mathbf{x} | k_i) = \frac{1}{(2\pi)^{d/2} \sigma^d} \cdot \frac{1}{M_i} \sum_{j=1}^{M_i} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{x} - \mathbf{c}_{ij})^T (\mathbf{x} - \mathbf{c}_{ij})\right), \quad i = 1, \dots, K \quad (2.1)$$

където $\mathbf{c}_{ij}$ е j -тия обучаващ вектор от клас k_i , $\mathbf{x}$ е входния тестов вектор, d е размерността на вектора, M_i е броя обучаващи вектори в клас k_i , а σ е стандартното отклонение

Исходния слой реализира правилото на Бейс (2.2) за класификация с минимален риск [106]:

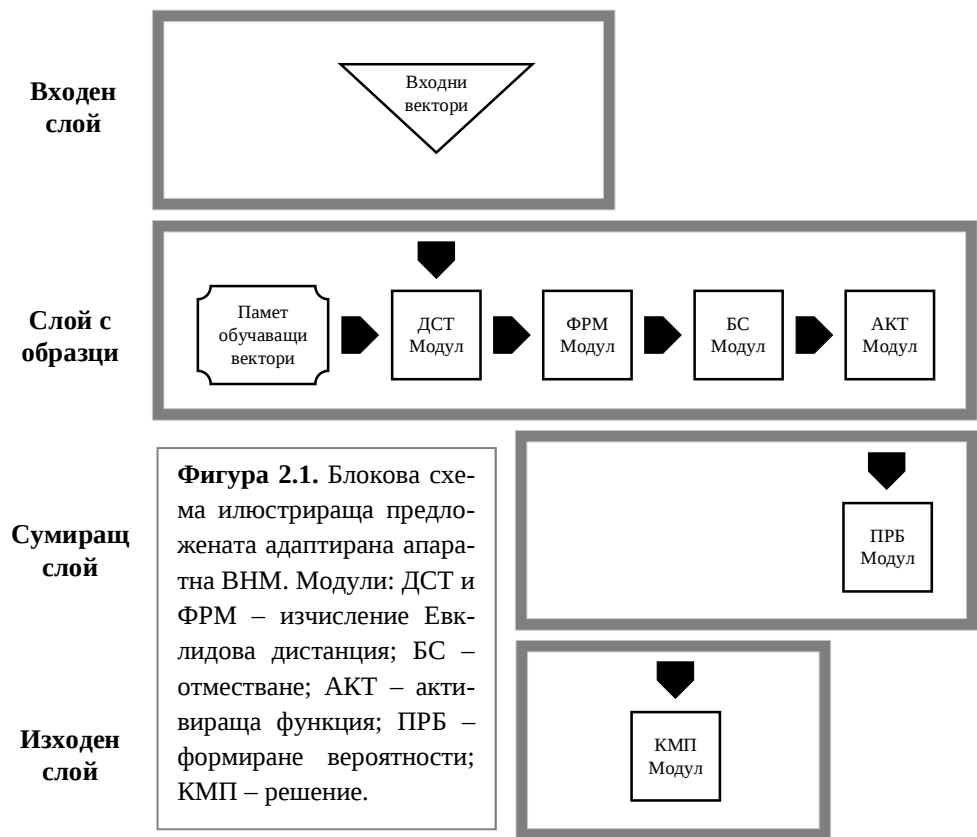
$$D(\mathbf{x}) = \underset{i}{\operatorname{argmax}} \{P(k_i)p_i(\mathbf{x} | k_i)\}, \quad i = 1, \dots, K \quad (2.2)$$

2.2. Апаратна реализация на ВНМ

Реализираният FPGA дизайн на ВНМ е повлиян от три важни конструктивни съображения: използвани ресурси на чипа, класификационна точност и изчислителна скорост. Представянето на данните в дизайна е посредством фиксирана запетая, с цел оптимизиране на бързодействието и използваната площ на чипа. Разрядността на данните с фиксирана запетая във всеки един от модулите е определена в съответствие с критериите на даденото приложение на модула [П.1].

2.2.1. Модел на апаратна ВНМ

На Фигура 2.1 е представена блокова схема илюстрираща потока от данни, както и основните модули в предложената FPGA реализация на изцяло паралелен дизайн на ВНМ класификатора. Разработеният FPGA дизайн на ВНМ е насочен към обработката на бинарни изображения [П.1].



2.2.1.1. Моделиране на първия слой на ВНМ

При апаратната реализация входния вектор се приема паралелно, т.е. всеки елемент от вектора е реализиран като вход за невронната мрежа.

2.2.1.2. Моделиране на втория слой на ВНМ

В случаите когато се работи с бинарни изображения, квадратичната операция при изчислението на Евклидовата дистанция в (2.1), се заменя с абсолютна стойност, с което се елиминира използването на множители и др. сложни структури. Това намалява използваната площ от чипа без да редуцира класификационната точност.

Формирането на изходите на отместващия модул (БС модул) в предложения паралелен дизайн, е директен резултат от едновременното умножение на входните за модула данни по предварително изчислен коефициент (*bias*) на база на изглаждащия фактор σ :

$$bias = \frac{\sqrt{-\log(0.5)}}{s} \quad (2.3)$$

С цел редуциране на използваните ресурси е разработен вариант на ВНМ без отместващ модул. Наблюдава се незначителна деградация в класификационната точност, която за конкретното приложение е определена като приемлива. В Секция 2.3 е разгледана реализацията на този модул и свързаните с него предизвикателства.

За изхода на ФРМ модула е определено, че разрядност на данните от пет бита е достатъчна за поддържане на класификационната точност.

Експоненциалната функция е реализирана като таблица на съответствия (LUT) попълнена с 8 битови стойности с фиксирана запетая. Тези стойности са интерпретирани като стойности с дву-битова цяла част и шест-битова дробна част.

2.2.1.3. Моделиране на третия слой на ВНМ

В ПРБ модула задаването на съответните класове от интерес за вече изчислените неврони е извършено чрез фиксирано кодиране. Това е с оглед целите на поставената задача за оптимално използване на наличните ресурси.

Реализацията на операция деление в ПРБ модула във FPGA чипа се оказва предизвикателство, не само поради естеството на операцията, но и поради изискванията към представянето на данните, както и желаната производителност. В текущият случай операцията деление е реализирана с помощта на принципа за деление на Goldschmidt [107].

2.2.1.4. Моделиране на изходния слой на ВНМ

Изходен или КМП модул реализира правилото на Бейс за класификация с минимален риск (2.2), като за неговата реализация тук е предложено изпълнение с компаратор.

2.2.2. Експериментална постановка

Разгледана и оценена е функционалността на три FPGA реализации на ВНМ с различна сложност. Апаратните и програмните реализации на ВНМ са изпитани в общ експериментален протокол, използващ публично достъпната база данни от бинарни изображения SPECT Heart Data Set. В проведеното сравнително оценяване са използвани 20, 50 и 80 обучаващи вектора за обучението на програмната и FPGA базираната ВНМ.

2.2.3. Резултати

Таблица 2.1. Грешки от класификация за реализациите на ВНМ - 20, 50 и 80 неврона.

Брой неврони	Програмна (грешка в %)	Апаратна (грешка в %)
20 неврона	39.0%	40.1%
50 неврона	30.5%	32.6%
80 неврона	29.4%	30.0%

Таблица 2.2. Използвани ресурси за реализацията на трите ВНМ във FPGA – 20, 50 и 80 неврона.

Сложност	Общо ЛЕ	Изчислително време (тактове)
20 неврона	13 700 ЛЕ	3 такта
50 неврона	26 413 ЛЕ	10 такта
80 неврона	43 255 ЛЕ	7 такта

Таблица 2.3. Използвани ресурси за всеки от основните модули на реализираните дизайни – 20, 50 и 80 неврона.

Модул	20 неврона (брой ЛЕ)	50 неврона (брой ЛЕ)	80 неврона (брой ЛЕ)
ДСТ Модул	12 107	22 458	37 162
ФРМ Модул	943	2 581	4 197
АКТ Модул	260	700	1 120
ПРБ Модул	367	628	742
КМП Модул	21	32	32

2.3. Оптимизиране на апаратния модел на ВНМ

Поради естеството на данните е възможно допълнително редуциране на използваните ресурси чрез замяната на аритметичното действие изваждане с логическо изключващо ИЛИ без промени в крайния резултат. Тази модификация води да значително намаляване на заетите ресурси за дадения модул.

Друга възможност за редуциране на необходимите ресурси се открива при реализацията на отместването или БС модула. По-точно възможността за реализацията на невронната мрежа без въпросния модул. Практически липсата на реализиран модул за отместване води до фиксирана стойност на коефициента на изглаждане. Тъй като се приема, че отместването в този случай ще бъде равно на единица, а самото отместване се изчислява на база стойността на изглаждащия фактор σ (2.3). Отсъствието на такъв модул може да влоши класификационните качества на невронната мрежа (поради невъзможността за настройка на полето на функцията на разпределение на вероятността), но правилен подбор на обучаващите данни позволява това да се компенсира до голяма степен.

Експонентата в предложения дизайн е реализирана като таблица на съответствия. Входните адреси за която са формирани от предходния модул, което в текущата реализация може да бъде ФРМ модула или БС модула. Косвено последствие от реализирането на модул за отместване е необходимостта от нарастване на таблицата на съответствия за покриване на възможните получени адреси, тъй като изхода на отместващия модул се явява вход на модула реализиращ експоненциалната функция.

Решение на проблема е намерено в реализиране на адресен дешифратор. Добавената допълнителна логика комбинира определен диапазон от стойности изчислени в отместващия модул като единичен адрес за таблицата на съответствия.

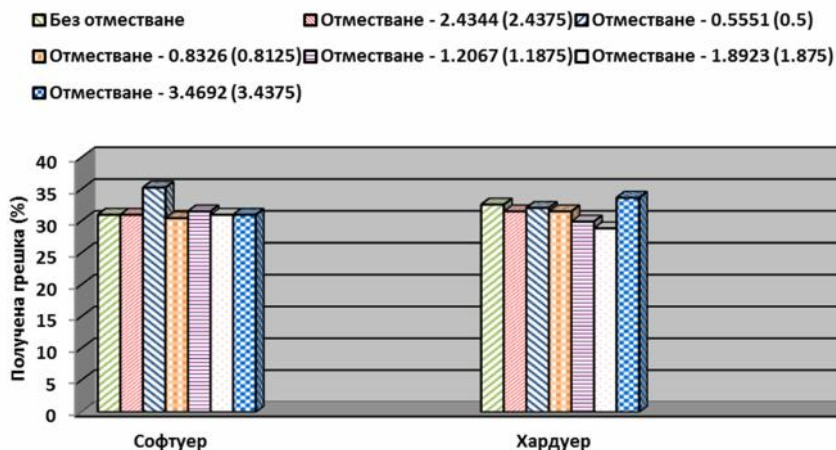
2.3.1. Експериментална постановка

Разгледана и оценена е функционалността на FPGA базирани реализации на ВНМ със сложност от 50 неврона. Оценката е извършена с помощта на развойната платформа DE2-115, използваща Altera Cyclone IV. FPGA дизайн е разработен в средата Quartus II 14.0, чрез език за описание на хардуер Verilog HDL и верифициран в ModelSim 10.1e.

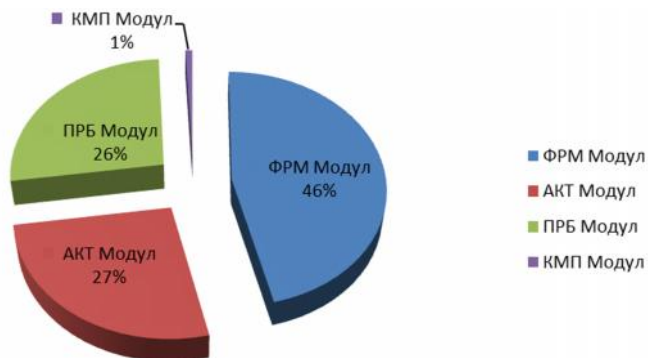
Апаратните и програмните реализации на ВНМ са изпитани в общ експериментален протокол, използващ публично достъпната база данни от бинарни изображения SPECT Heart Data Set.

В проведеното сравнително оценяване са използвани 50 обучаващи вектора за обучението на програмната и FPGA базираната ВНМ. Броят обучаващи вектори е получен чрез подбор на 50 от наличните 80 обучаващи вектора. Във всички експерименти броят тестови вектори е 187, класифицирани като нормални или аномални.

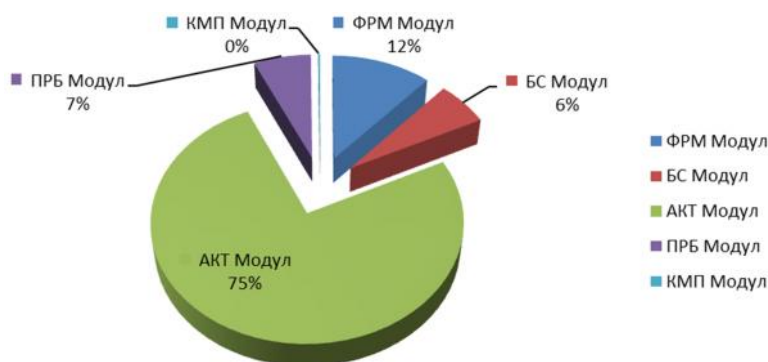
2.3.2. Резултати



Фигура 2.2. Грешка при класификация за различни стойности на изглаждащия фактор σ .



Фигура 2.3. Разпределение на ресурсите без реализирано отнемстване – 2485 ЛЕ.



Фигура 2.4. Разпределение на ресурсите с реализирано отнемстване – 9906 ЛЕ.

За реализиране на мрежа без отнемстващ модул при програмния вариант е използван изглаждащ фактор σ със стойност ($\sigma = 0.8299$) еквивалентен (2.3) на отнемстване 1.0032, т.е. приблизително 1. По същия начин са изчислени и други стойности на отнемстването за даден изглаждащ фактор σ . Поради физическите ограничения на чипа стойностите на отнемстването в апаратния вариант са представени с прецизност от 6 бита.

На Фигура 2.3 и Фигура 2.4 е показано разпределението на броя логически елементи използвани за основните модули в реализирания дизайн с Quartus II 15.0.2. ДСТ модула, както и модула на най-горното ниво на дизайна не са показани, тъй като те съдържат логика, използваща незначителна част ресурси.

2.4. Оптимизационни стъпки при апаратната реализация на ВНМ

За случая на ВНМ за класификация на бинарни изображения е предложена методика от шест стъпки за оптимизация на използваните FPGA ресурси. Основно изискване за успешно прилагане на предложените стъпки е входните вектори за

класификация, както и обучаващите данни, да се състоят от еднобитови признаци. Броят признаци, както и броя обучаващи данни е без значение. Косвено ограничение възниква от капацитета на избрания FPGA чип. Шестте главни стъпки са както следва:

- i. *Опростяване на изчисленията свързани с тегловните коефициенти до логически операции.* – Изчисляването на Евклидовата дистанция (2.1) се заменя с логическо изключващо ИЛИ.
- ii. *Редуциране броя на аритметичните операции умножение.* – Премахване на отместването и внимателно подбиране на обучаващите данни.
- iii. *Използване на таблица на съответствия за реализирането на нелинейни операции.* – Всички възможни адреси постъпващи към таблицата на съответствия да бъдат покрити.
- iv. *Оптимизиране на таблицата на съответствия.* – Оптимизацията на таблицата на съответствия се състои от три действия, които се свеждат до обединяване на определен брой адреси. За правилната работа на това решение е необходима реализацията на адресен дешифратор.
- v. *Използване на гъвкав алгоритъм за деление.* – Позволява допълнителна оптимизация на невронната мрежа за бързодействие или точност.
- vi. *Реализиране на изходния слой.* – Използване на компаратор.

2.5. Обобщения и изводи

- Разработен е приложно-оптимизиран паралелен апаратен модел на ВНМ, подходящ за бърза обработка на бинарни изображения [П.1], [П.3].
- Оценени са три FPGA реализации с различна сложност, по критерии използвани ресурси, класификационна точност, и скорост на изчисленията [П.1].
- Експериментално е проверена пригодността на оптимизирания паралелен апаратен модел на ВНМ за класификацията на бинарни изображения [П.1], [П.5].
- Предложени са промени в алгоритъма на ВНМ, водещи до намаляване площта на апаратната реализация във FPGA чипа при незначително влошаване (1-3%) на точността на класификация [П.5].
- Създадена е методика за оптимизиране на използваните FPGA ресурси при реализацията на апаратен модел на ВНМ за бинарни изображения [П.1], [П.5].

2.6. Приноси към Глава 2

- Създадена е методика, позволяваща оптимизиране на използваните FPGA ресурси при реализацията на апаратен модел на вероятностната невронна мрежа за бинарни изображения.
- Разработен е приложно-оптимизиран паралелен апаратен модел на вероятностната невронна мрежа подходящ за бърза класификация на бинарни изображения.

Глава 3: Активиращи функции на вероятностната невронна мрежа в електрически програмируеми матрични кристали

В Глава 3 е представен и изследван адаптиран метод за апроксимация на експоненциалната активираща функция при апаратна реализация на ВНМ. Също така са изследвани и оценени активиращи функции, подходящи при апаратна реализация на ВНМ, които да заместят експоненциалната активираща функция.

3.1. Активиращи функции за вероятностната невронна мрежа

Като алтернатива на традиционната активираща функция (2.1), са разгледани други възможности за моделиране на функцията на разпределението на вероятността [105], [108], [109], [14]:

$$f(x) = \frac{1}{n(2l)^p} \sum_{i=1}^n 1, \quad \text{when all } |x_i - c_{ij}| \leq l \quad (3.1)$$

$$f(x) = \frac{1}{n l^p} \sum_{i=1}^n \prod_{j=1}^p \left[1 - \frac{|x_i - c_{ij}|}{l} \right], \quad \text{when all } |x_i - c_{ij}| \leq l \quad (3.2)$$

$$f(x) = \frac{1}{n(p l)^p} \sum_{i=1}^n \prod_{j=1}^p \left[1 + \frac{(x_i - c_{ij})^2}{l^2} \right]^{-1} \quad (3.3)$$

3.2. Апроксимация на активиращата функция на ВНМ

Както е описано в Глава 2, експоненциалната функция е реализирана чрез таблица на съответствия, предоставяйки един от най-бързите начини за изчисление. Необходимостта от запазване на кратко изчислително време възпрепятства използването на единична таблица на съответствия, която да е споделена измежду всички неврони [57]. За настройка на таблицата на съответствия са изследвани ефектите при различни нейни конфигурации върху производителността на ВНМ и изразходваните апаратни ресурси [П.2].

Нека бинарно изображение е представено с 22 признака с прецизност 1 бит, тогава, таблицата на съответствия ще е с дълбочина 23 адреса [0,22]. Създаването на таблица на съответствия със стъпка единица и избраната дължина на думата не позволява да се запълнят всички адреси, което ще влоши точността. При увеличаване на прецизността (броя битове) за представяне на дробната част се увеличават необходимите ресурси за конструиране на желаната таблица на съответствия, както и изискванията към логиката която ще обработва изхода ѝ.

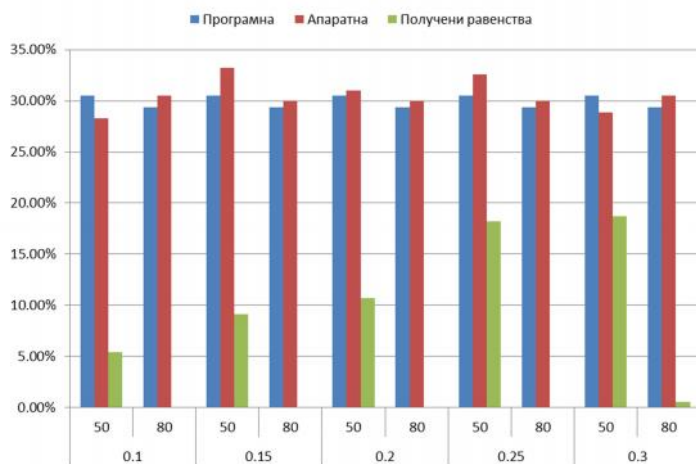
За преодоляване на този проблем е използвана по-малка стъпка за формиране на таблицата [П.2].

3.2.1. Експериментална постановка

За проверка поведението на ВНМ при различните апроксимации на активиращата функция, са реализирани две FPGA-базирани реализации с различна сложност. Апаратните и програмните реализации на ВНМ са изпитани в общ експериментален протокол, използващ публично достъпната база данни от бинарни изображения SPECT Heart Data Set. В проведеното сравнително оценяване са използвани 50 и 80 обучаващи вектора. Всяка от реализациите е оценена за пет различни таблици на съответствия, реализирани със стъпка 0.10, 0.15, 0.20, 0.25 и 0.30.

3.2.2. Резултати

Реализацията с 80 неврона показва устойчиво поведение дори при таблицата на съответствия съставена с най-грубата стъпка между стойностите. Стойностите на грешката не се изменят значително и се избягват случаите на равенство на вероятностите за принадлежност към клас. Това показва добрите качества и устойчивост на ВНМ при манипулации целящи опростяване от гледна точка на транслиране на архитектурата в апаратна среда като FPGA.



Фигура 3.1. Грешка от класификация с ВНМ от 50 и 80 неврона при различен размер на таблицата на съответствия (показана е грешката в проценти).

3.3. Алтернативни активиращи функции за ВНМ

Активиращи функции (3.1)-(3.3) използват математически операции които са удобни за апаратна реализация във FPGA. За оценяване на тяхната приложимост, е представено сравнение с традиционният модел (2.1). За удобство, изследваните активиращи функции са означени като AF-1 (2.1), AF-2 (3.1), AF-3 (3.2), и AF-4 (3.5) [П.6].

3.3.1. Експериментална постановка

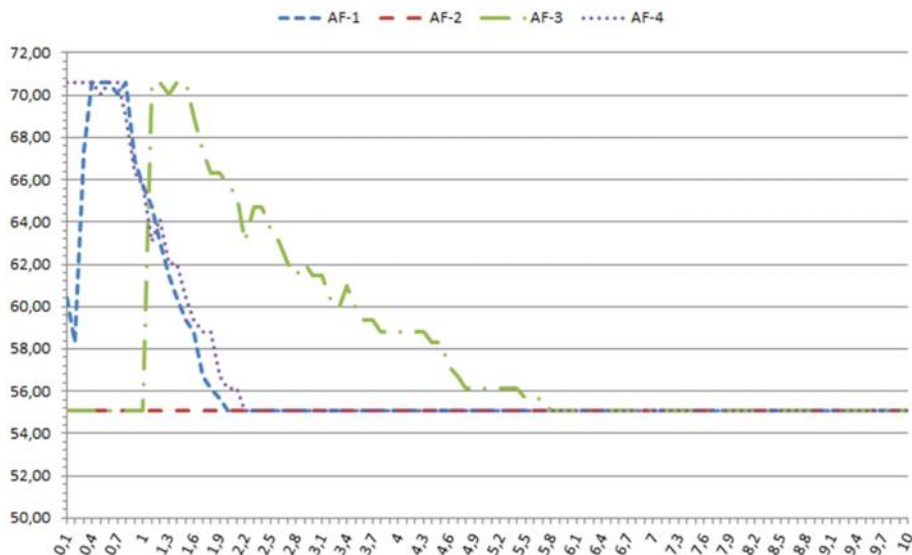
Трите избрани активиращи функции (3.1)-(3.3) са сравнени с традиционния модел (2.1) при класификация на бинарни изображения – база данни SPECT Heart Data Set.

Разгледани са три случая с различен брой неврони на ВНМ за всеки от оценителите на функцията на разпределението на вероятността. Използвани са 80 обучаващи вектора, както и подгрупи от 50 и 20 обучаващи вектора. По-ниският брой неврони (обучаващи вектори) допринася за намаляване размера на използваните ресурси от FPGA чипа, или използване на по-малък FPGA чип, което от своя страна влияе на крайната себестойност на изделието и на общото потребление на енергия.

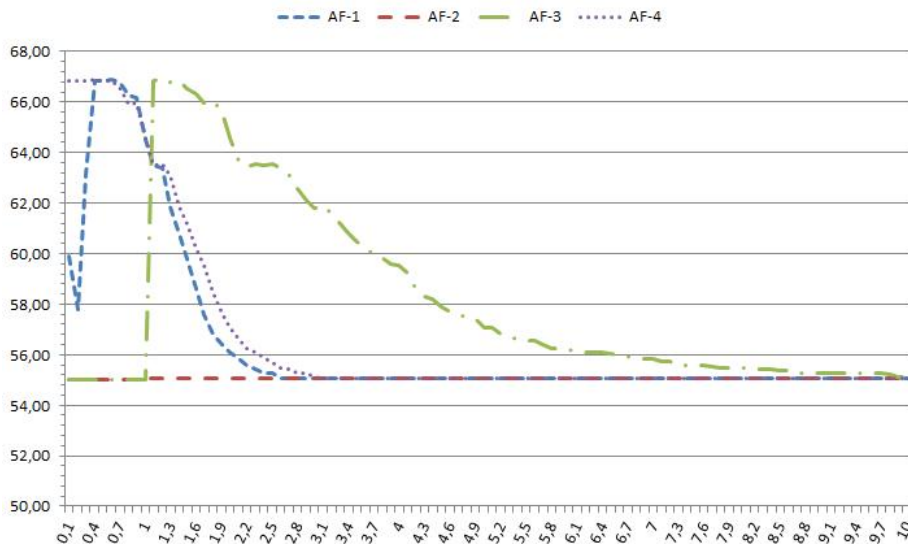
3.3.2. Резултати

Експериментите са извършени за 100 различни стойности на двата коефициента σ и λ , в диапазона от 0.1 до 10, със стъпка от 0.1. В проведените експерименти не е търсена оптимална стойност на параметрите, а е изследвана прецизността на изследваните модели при еднакви условия. Във всеки от проведените експерименти стойността на коефициента σ и коефициента λ си съответстват.

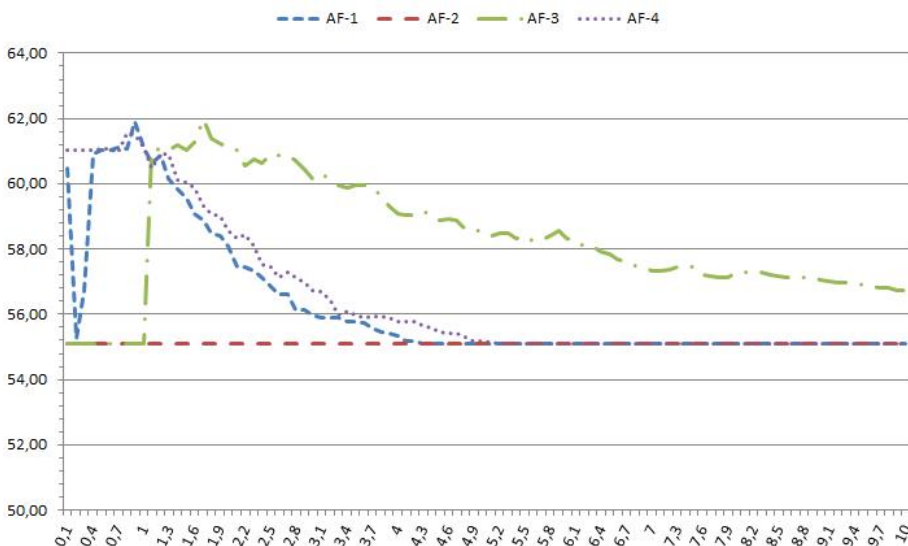
Реализациите с AF-3 и AF-4 имат значително по-добро поведение от колкото AF-2, върху текущия проблем. При случаите в които ВНМ е реализирана с функциите AF-3 и AF-4 се наблюдават сходни резултати до тези на традиционната активираща функция за ВНМ.



Фигура 3.2. ВНМ с 80 невронна в слоя с образци (абсцисата показва стойностите на коефициентите σ и λ , а ординатата точността в проценти).



Фигура 3.3. ВНМ с 50 невронна в слоя с образци (абсцисата показва стойностите на коефициентите σ и λ , а ординатата точността в проценти).



Фигура 3.4. ВНМ с 20 невронна в слоя с образци (абсцисата показва стойностите на коефициентите σ и λ , а ординатата точността в проценти).

3.4. Обобщения и изводи

- Предложен е адаптиран метод за апроксимация на експоненциалната активираща функция при апаратна реализация на ВНМ, предоставящ приемлив компромис между ресурси (необходима логика), изчислителна скорост и точност [П.2].
- Изследвани и оценени като подходящи за апаратна реализация са две активиращи функции (AF-3 и AF-4) на ВНМ, показващи сходна точност с експоненциалната активираща функция при класификация на бинарни изображения [П.6].

3.5. Приноси към Глава 3

- Подобрен е метод за апроксимация на експоненциалната активираща функция при апаратна реализация на вероятностната невронна мрежа.
- Предложени, изследвани и оценени са алтернативи на експоненциалната активираща функция, подходящи при апаратна реализация на вероятностната невронна мрежа.

Глава 4: Сравнителен анализ за стратегии за обучение на вероятностната невронна мрежа

В Глава 4 са изследвани и оценени деветнадесет конфигурации за обучение на ВНМ, използващи различни настройки на алгоритмите диференциална еволюция и оптимизация с рояк от частици. Избран е метод за обучение на ВНМ, удобен за апаратна реализация във FPGA.

4.1. Обучение на вероятностната невронна мрежа

Първата стъпка от обучителния процес на ВНМ създава първия скрит слой (слой с образци), който запомня обучаващите вектори. Втората стъпка от обучението включва определяне на изглаждащия фактор σ .

4.2. Диференциална еволюция

Диференциална еволюция (DE) е добре познат итерационен оптимизационен метод, който систематично търси в пространство с възможни решения чрез популация от частици. При всяка итерация, наречена поколение, се извършват три стъпки, наречени – мутация, рекомбинация и избор [111].

4.3. Оптимизация с рояци от частици

Оптимизация с рояци от частици (PSO) е стохастичен оптимизационен метод, базиран на популация от частици. След инициализацията (на случаен принцип) на началните стойности на всички частици в d -мерното пространство, техните нови позиции са преизчислени итерационно, чрез изчисление на скоростта и опресняване на текущата им позиция [113], [114].

4.4. Експериментална постановка

Експерименталното оценяване е проведено чрез използване на базата данни Parkinson Speech Dataset with Multiple Types of Sound Recordings. Разгледан е класификатор в два класа, тестван по схема с десетократно кръстосано валидиране, при разпределение на данните 90% за обучение и 10% за тестове. Изследвана е целесъобразността на деветнадесет конфигурации, спрямо тяхната приложимост за определянето на изглаждащия фактор σ в хетероскедастична ВНМ. Всички конфигурации са базирани на различни версии на DE и PSO алгоритми. По-конкретно сред тях са: (а) десет стратегии за DE, с мутационни оператори базирани на best/1/exp, rand/1/exp, rand-to-best/1/exp, best/2/exp, rand/2/exp, best/1/bin, rand/1/bin, rand-to-best/1/bin, best/2/bin и rand/2/bin, както и (б) девет PSO алгоритъма, сред които Common PSO w/inertia, Trelea type 1, Trelea types 2, Clerc's Constricted PSO – Type 1, UPSO с пет различни унифициращи фактора ($u=0$, $u=0.25$, $u=0.5$, $u=0.75$ и $u=1$). При сравнителното оценяване са използвани две популации съответно с 20 и 50 частици за всички обучаващи конфигурации.

Таблица 4.1. Получена осреднена грешка от проведените експерименти.

Оптимизационен алгоритъм	Популация с размер 20 (грешка в %)	Популация с размер 50 (грешка в %)
(1) DE/best/1/exp	25.8%	25.3%
(2) DE/rand/1/exp	26.9%	25.9%
(3) DE/rand-to-best/1/exp	25.8%	25.4%
(4) DE/best/2/exp	27.2%	26.8%
(5) DE/rand/2/exp	28.8%	27.6%
(6) DE/best/1/bin	25.8%	25.1%
(7) DE/rand/1/bin	26.1%	25.8%
(8) DE/rand-to-best/1/bin	25.7%	25.4%
(9) DE/best/2/bin	27.3%	26.8%
(10) DE/rand/2/bin	28.1%	27.5%
(11) Common PSO	28.2%	27.2%
(12) Trelea type 1	28.7%	27.8%
(13) Trelea type 2	28.6%	27.6%
(14) Clerc's Constricted PSO - Type 1"	28.1%	27.6%
(15) UPSO – $u=0$	25.8%	25.3%
(16) UPSO – $u=0.25$	25.5%	25.0%
(17) UPSO – $u=0.5$	25.8%	25.0%
(18) UPSO – $u=0.75$	26.1%	25.3%
(19) UPSO – $u=1$	26.3%	25.7%

4.5. Резултати

В Таблица 4.1 е представена осреднената грешка при класификация (в проценти), изчислена след петдесет последователни експеримента за всяка от конфигурациите, както и за всеки размер на популацията.

4.5.1. Статистически анализ на резултатите

За изследване на статистическата разлика между резултатите за отделните конфигурации е използван статистически анализ по двойки. Получаване на заключение относно повече от едно сравнително изследване увеличава риска от натрупване на грешки, в следствие на Family-Wise Error Rate (FWER) [119], поради което за откриване на възможните разлики между изследваните конфигурации е използван мултиплициран тест за сравнения на Friedman ($1 \times N$) и последващ анализ на Hochberg. Процедурата е използвана и за двете популации. За контролна конфигурация и при двете популации е използвана (16) UPSO $u=0.25$.

Проведен е и многократен тест на Friedman ($N \times N$) с последващ анализ на Holm за всички конфигурации. Оценени са 171 хипотези за сходство сред деветнадесетте конфигурации. Оценяването е извършено и за двете популации.

Статистическите тестове не успяват да отхвърлят хипотезата за сходство между следните конфигурации: (1) DE/best/1/exp, (6) DE/best/1/bin, (16) UPSO $u=0.25$ и (17) UPSO $u=0.5$.

4.5.2. Разлики между популациите

Алгоритъм (16) UPSO – $u=0.25$ се откроява като алгоритъма постигащ най-ниска грешка представена в проценти сред разгледаните конфигурации и за двете популации (20 и 50 частици).

Експериментите са проведени следвайки експерименталният протокол описан в Секция 1.5.2, само за конфигурациите: (1) DE/best/1/exp, (6) DE/best/1/bin, (16) UPSO $u=0.25$ и (17) UPSO $u=0.5$.

Извършен е многократен тест на Friedman с последващ анализ на Holm на резултатите между различните популации (10, 20, 30, 40 и 50 частици) на всяка една от конфигурациите. От резултатите може да се заключи, че общо тридесет и три двойки се различават при ниво на значимост от $\alpha=0.01$, тридесет и осем при $\alpha=0.05$ и тридесет и девет при $\alpha=0.1$.

Проведени са статистически тестове из между четирите конфигурации за всяка от разгледаните популации. При повечето случаи, не е открита статистическа разлика дори при ниво на значимост от $\alpha=0.1$. Единствено по-забележими са два случая при големина на популацията от 40 частици, където стойността на p е близо до ниво на значимост от $\alpha=0.01$.

4.5.3. Изчислителна сложност и компромиси

За разглеждането на апаратна реализация във FPGA (например с Altera Cyclone IV), се оценяват апаратните ресурси, както и е необходима оценка за сложността на всяка от реализациите.

Разглеждайки цялостната сложност на избраните конфигурации, се забелязва, че UPSO алгоритъма би изисквал повече апаратни ресурси за реализация – поради наличието на комбинация на локален и глобален подход на търсене.

Въз основа на дискутираните съображения и получени сходни резултати при класификация, методът за обучение означен като DE/best/1/bin се разглежда като подходящ за реализация във FPGA, спрямо конфигурациите UPSO и DE/best/1/exp, особено когато се вземе предвид обема на изискваните ресурси.

4.6. Обобщения и изводи

- Обучаващи стратегии показващи най-добри резултати за изследвания проблем и определени като най-подходящи са локалните обучаващи стратегии - local PSO, DE/best/1 [П.4].
- Като оптимизационен алгоритъм постигащ най-добри резултати, от гледна точка на точност при обучението на BHM е определен Unified PSO (UPSO) с унифициращ фактор от $u=0.25$ [П.10].
- Конфигурацията, базирана на стратегията DE/best/1/bin, е определена като най-удачния кандидат за апаратна реализация във FPGA чип [П.10].

4.7. Приноси към Глава 4

- Дефинирани са критерии за удобство при апаратна реализация с FPGA и е избран метод за обучение на вероятностната невронна мрежа.

Глава 5: Локално рекурентна вероятностна невронна мрежа за електроенцефалографски сигнали в електрически програмируеми матрични кристали

В Глава 5 е разработен паралелен апаратен модел на локално рекурентната вероятностна невронна мрежа (ЛРВНМ) с висока разрядност на предавателната функция в рекурентния слой. Апаратният модел е валидиран чрез сравнително експериментално изследване с класификатор с опорни вектори (SVM), както и с програмна реализация на ЛРВНМ в среда на MATLAB. Предложена е нова разновидност на ЛРВНМ, при която локално рекурентния слой използва неврони с активираща функция от семейството на ReLU.

5.1. Архитектура на локално рекурентната вероятностна невронна мрежа

ЛРВНМ е изградена чрез вмъкването на допълнителен скрит слой, наречен рекурентен слой, между сумиращия слой и изходния слой на ВНМ [123], [106]. Както и при ВНМ за изчисление на вероятността за даден клас са използвани прозорци на Парзен реализирани чрез сума от Гаусови функции [106].

$$f_i(\mathbf{x}_p) = \frac{1}{(2\pi)^{d/2} \mathbf{S}_i^d} \cdot \frac{1}{M_i} \sum_{j=1}^{M_i} \exp\left(-\frac{1}{2\mathbf{S}_i^2} (\mathbf{x}_p - \mathbf{c}_{ij})^T (\mathbf{x}_p - \mathbf{c}_{ij})\right), \quad i=1, \dots, K. \quad (5.1)$$

Изходът от сумиращите неврони в локално рекурентния слой е [123]:

$$y_i(\mathbf{x}_p) = \left[b_{i,i} f_i(\mathbf{x}_p) - \sum_{\substack{k=1 \\ i \neq k}}^K b_{i,k} f_k(\mathbf{x}_p) \right] + \sum_{t=1}^N \left[a_{i,i,t} \bar{y}_i(\mathbf{x}_{p-t}) - \sum_{\substack{k=1 \\ i \neq k}}^K a_{i,k,t} \bar{y}_k(\mathbf{x}_{p-t}) \right], \quad i=1, \dots, K \quad (5.2)$$

където $f_i(\mathbf{x}_p)$ е функцията на разпределение на вероятността за всеки определен клас (k_i), $\mathbf{x}_p$ – входния вектор, K – броя класове, N – рекурентната дълбочина, $y_i(\mathbf{x}_{p-t})$ е нормализираната предишна изходна стойност за клас k_i , която е била задържана с t време, $a_{i,j,t}$ и $b_{i,j}$ са тегловни коефициенти.

Изходът на сумиращия неврон $y_i(\mathbf{x}_p)$ в рекурентния слой се подлага на нормализация, която запазва вероятностната интерпретация на изхода на рекурентния слой (sgm – сигмоидална активираща функция) [123]:

$$\bar{y}_i(\mathbf{x}_p) = \frac{sgm(y_i(\mathbf{x}_p))}{\sum_{j=1}^K sgm(y_j(\mathbf{x}_p))}, \quad i=1, 2, \dots, K \quad (5.3)$$

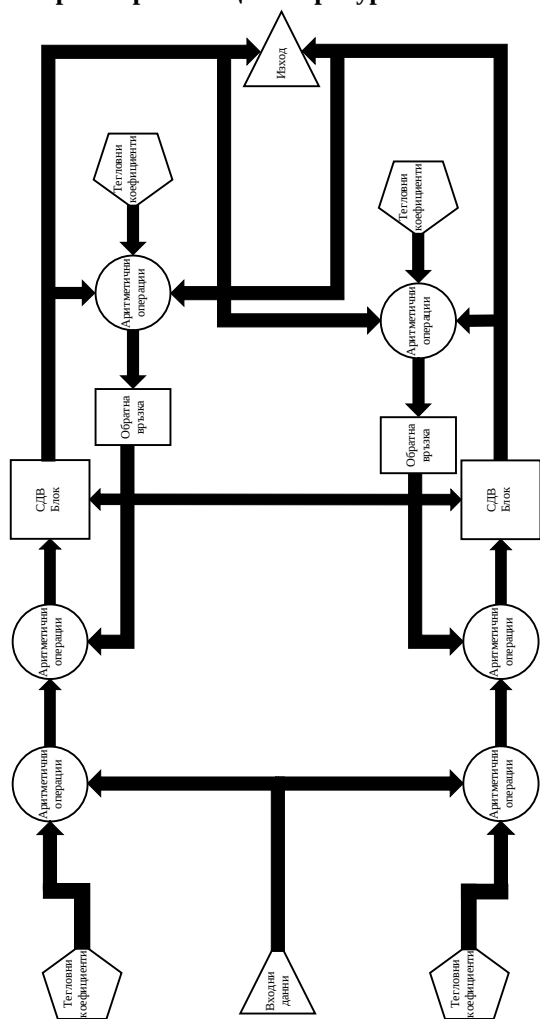
5.1.1. Обучение на ЛРВНМ

Обучението на ЛРВНМ протича в три стъпки. Първата стъпка създава топологията на мрежата и е аналогична с тази при обучението на ВНМ. Втората стъпка се състои в изчислението на изглаждащия параметър σ_i за всеки един от класовете или общо за всички класове (виж Глава 4) [124], [125], [126], [123]. Третата стъпка е свързана с изчислението на тегловните коефициенти на локално рекурентния слой, което включва минимизиране на функция на грешката [121], [122], [123]:

$$E(w) = \sum_{i=1}^K m_i (1 - P(k_i | X_{k_i})) P(k_i) + \frac{1}{K-1} \sum_{i=1}^K \sum_{\substack{j=1 \\ j \neq i}}^K m_j P(k_j | X_{k_i}) P(k_j), \quad (5.4)$$

За изчисление на тегловните коефициенти е използван оптимизационен алгоритъм, който постига задоволителни резултати при приемливо изчислително време.

5.2. Апаратна реализация на рекурентния слой

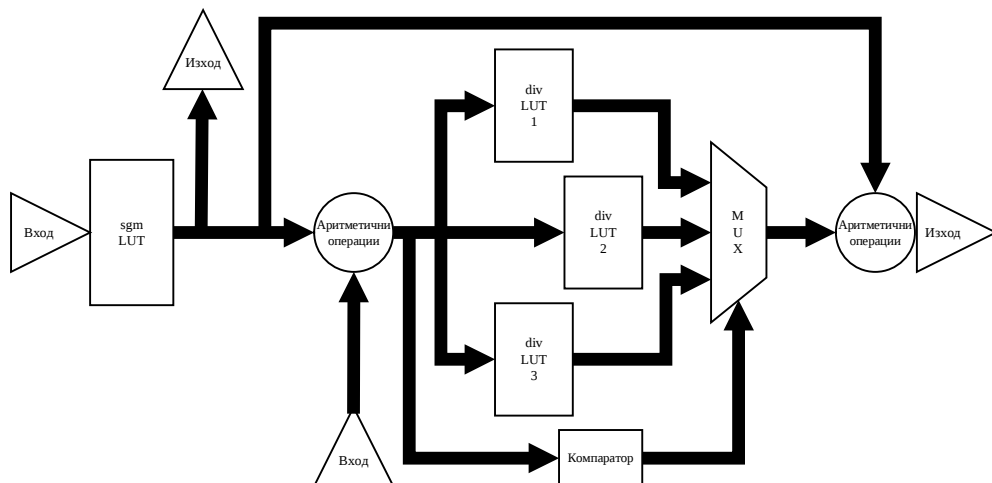


Фигура 5.1. Блокова схема на апаратната реализация на рекурентния слой в ЛРВНМ с дълбочина на рекурентност единица.

Апаратната реализация на рекурентния слой следва дефиницията на невроните от рекурентния слой в (5.2) и (5.3) [П.8].

Дизайнът е реализиран чрез представяне на данните във формат с фиксирана запетая, като в по-голямата част от дизайна дължината на думата е 17 бита. Най-старшият бит е запазен за знака, 5 бита са използвани за цялата част и 11 за дробната. Стойностите на входовете на рекурентния слой достигат максимална стойност на десетичното число 1, но стойностите на тегловните коефициенти на рекурентния слой не са ограничени до определен обхват. За избягване на този недостатък е

направен компромис за ограничаване на диапазона на претърсване за възможни решения за тегловните коефициенти до $[-15 + 15]$.



Фигура 5.2. Блокова схема на СДВ Блок.

СДВ блоковете съставляват логиката за сигмоидалната предавателна функция, както и логиката за операцията деление. Тези модули съдържат логика реализираща операции между числа представени с до 91 бита, съответно 46 бита за цялата част и 45 бита за дробната част. Необходимостта от такава висока прецизност в някои части на дизайна идва от наличието на обратна връзка от изходите на невроните в слоя. Аритметичните операции за предавателната функция и извършване на операция деление са реализирани като таблици на съответствия (Фигура 5.2). Използваната разрядност за реализацията на предавателната функция за дробната част е 45 бита, докато за цялата част е само един бит.

5.2.1. Експериментална постановка

Експериментите са проведени върху базата данни DEAP. Използвана е изчислената нормализирана енергия от 32 канално ЕЕГ от данните на десет участника. Данните са разделени на три групи: **обучаващи** – за създаване на топологията на невронната мрежа; **развойни** – за изчисляване на изглаждащия фактор σ и изчисляване на теглата в рекурентния слой; **тестови** – за оценка на работата на невронната мрежа. Разпределението на данните в проценти е както следва: 20% обучаващи, 20% развойни и 60% тестови. Целта на изследването е разпознаването между два класа: (а) клас *негативни* емоционални състояния и (б) обобщен клас *други*, включващ неутрални и позитивни емоционални състояния.

Оптимизационният алгоритъм (DE) за обучение на рекурентния слой е използван както с неограничено нарастване на стойностите на частиците в популацията, така и с ограничение в диапазона на нарастване до [-15 +15].

Разгледан е и вариант със следното разпределение на данните: 20% обучаващи, 60% развойни и 20% тестови. Обучението на мрежата е извършено по същия начин, както и в предходния случай на разпределение на данните – с неограничено нарастване на стойностите на популацията и с ограничение на стойностите на частиците в популацията в диапазона от [-15 +15].

При апаратните експерименти, вече обработените данни от първите два слоя на невронната мрежа са използвани като входни данни за апаратната реализация на рекурентния слой. Тегловните коефициенти получени след обучителната процедура с ограничен диапазон на стойностите на популацията са пренесени в апаратната реализация.

5.2.2. Резултати

Работната честота на дизайна е 22.22 MHz с 45 ns тактов период. Въпреки, че работната честота е ниска, изхода се изчислява за един такт, като ниската работна честота спомага и за намаляване на консумацията на енергия. За същата задача, но изпълнена в програмен вариант на персонален компютър с Intel i7 (4700MQ) процесор с максимална работна честота от 2.4 GHz и 16 GB RAM (DDR3 – 1600 MHz) са необходими приблизително ≈ 30 μ s за изчисление на изход при даден вход, т.е. 667 пъти по-добро бързодействие, в полза на FPGA реализацията.

Таблица 5.1. Точност при класификация за програмни и апаратни реализации на рекурентния слой при разпределение на данните: 20% обучаващи, 20% развойни и 60% тестови.

DEAP Участник №	Програмна – неограничена (точност %)	Програмна – ограничена (точност %)	Апаратна – ограничена (точност %)
2	80.1	80.1	48.7
11	72.0	60.2	63.1
17	79.4	79.4	85.8
21	56.7	56.7	50.1
22	99.9	99.9	99.9
24	60.1	60.1	56.8
28	91.6	91.6	93.6
29	34.8	34.8	50.0
30	95.6	95.6	95.4
32	67.5	67.5	66.8
Осреднена	73.8	72.6	71.0

Таблица 5.2. Точност при класификация за програмни и апаратни реализации на рекурентния слой при разпределение на данните - 20% обучаващи, 60% развойни и 20% тестови.

DEAP Участник №	Програмна – неограничена (точност %)	Програмна – ограничена (точност %)	Апаратна – ограничена (точност %)
2	98.7	98.7	98.9
11	97.3	54.6	61.6
17	59.8	55.6	50.0
21	55.7	55.7	50.7
22	99.9	59.1	59.2
24	75.0	75.0	96.5
28	99.9	99.9	100.0
29	77.7	77.7	74.9
30	55.6	55.6	50.0
32	98.2	55.6	50.0
Осреднена	81.8	68.7	69.2

Разгледаните реализации (програмни и апаратни) на ЛРВНМ, показват по-добри резултати в сравнение с класификатор с опорни вектори, (докладвана средна точност от 65.7%) [128], [П.8].

5.3. Преосмисляне на невронния модел на ЛРВНМ

За активираща функция в рекурентния слой на ЛРВНМ се използва сигмоидалната активираща функция, която преобразува (5.5) независимата променлива x (реални числа със знак) в стойности в интервала $[0, 1]$ [121]:

$$f(x) = \frac{1}{(1 + \exp^{-x})} \quad (5.5)$$

Първата разгледана алтернатива на невронния модел със сигмоидална активираща функция е стандартна ReLU предавателна функция [136] :

$$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (5.6)$$

Следващата взета предвид активираща функция, която е и модификация на ReLU е Leaky ReLU [137] :

$$f(x) = \begin{cases} x, & x \geq 0 \\ const \cdot x, & x < 0 \end{cases} \quad (5.7)$$

където $const$ е фиксирана скаларна стойност, която пропуска отрицателни стойности с цел избягване на насищането в долната граница.

Третата разгледана алтернатива на сигмоидалната функция е отново модификация на ReLU, наречена Clipped ReLU [133]:

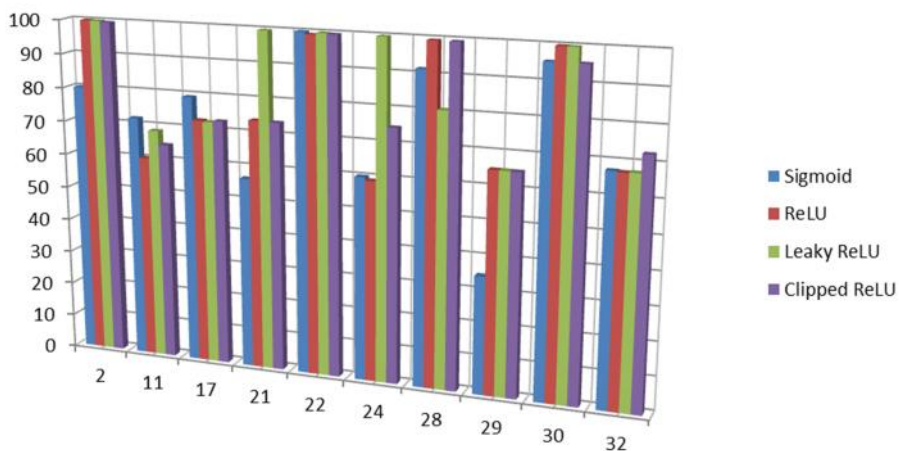
$$f(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x < const \\ const, & x \geq const \end{cases} \quad (5.8)$$

където *const* в този случай е максимална стойност, над която всички стойности превишаващи я биват приравнени на *const*.

5.3.1. Експериментална постановка

Експериментите са проведени за четири различни невронни модела на ЛРВНМ. Моделите са съответно с четирите различни активационни функции в рекурентния слой на мрежата, описани в Секция 5.3, а именно – (i) сигмоидална, (ii) ReLU, (iii) Leaky ReLU и (iv) Clipped ReLU. За проверка поведението на различните модели на неврони в рекурентния слой е използвана DEAP базата данни. Използваните данни са от десет участника, като класификацията се извършва въз основа на изчислената нормализирана кратковременна енергия от 32 канален ЕЕГ сигнал. Целта на изследването е разпознаването между два класа: (а) клас *негативни* емоционални състояния и (б) обобщен клас *други*, включващ неутрални и позитивни емоционални състояния. Използваните данни са разделени в три групи: обучаващи данни – 20%; развойни данни – 20%; тестови данни – 60%. Задоволителни резултати при обучението на рекурентния слой за текущия проблем е получен с оператор DE/rand-to-best/1/bin.

5.3.2. Резултати



Фигура 5.3. Обобщени резултати при използването на ReLU базирани предавателни функции и сигмоидална предавателна функция за ЛРВНМ (точност %).

Получените резултати потвърждават, че изследваните модели на рекурентен неврон с ReLU активираща функция подобряват прецизността на ЛРВНМ, спрямо сигмоидалния модел. Анализ на резултатите показва, че ЛРВНМ базирана на Leaky ReLU дава най-добра точност при класификация за разгледания проблем. Реализациите с Clipped ReLU и ReLU отстъпват по този критерий. Модификацията на ReLU – Leaky ReLU успява да надгради успеха на чистата ReLU функция поради възможността си да елиминира събитията на насищане в долната граница. Успехът на Clipped ReLU от своя страна се дължи на възможността за ограничаване на нарастването на стойностите в горната граница.

5.4. SatReLU активираща функция в рекурентния слой на ЛРВНМ

Предложена е нова активираща функция, наречена SatReLU, която да се използва в невроните от рекурентния слой на ЛРВНМ:

$$f(x) = \begin{cases} 0.001 \cdot const, & x < 0 \\ x, & 0 \leq x < const \\ const, & x \geq const \end{cases} \quad (5.9)$$

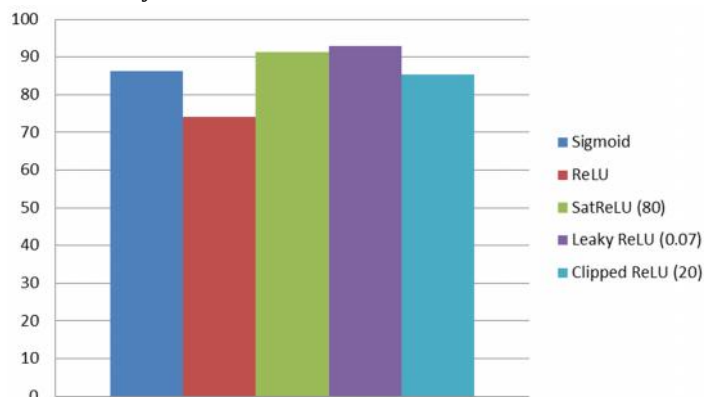
където *const* е максимална стойност на насищане, над която всички стойности биват приравнени на *const*. Стойността на *const* също се използва за дефиниране на праг за отрицателните стойности на входа, който обаче е значително по-малък по стойност.

Сравнението между (5.5) и (5.9) показва че SatReLU в значителна степен намалява изчислителната сложност и елиминира необходимостта от трудната за реализиране във FPGA сигмоидална активираща функция.

5.4.1. Експериментална постановка

Дефинирани са пет различни невронни модела на ЛРВНМ, които използват различни активиращи функции в рекурентния слой: (i) Sigmoid (сигмоидална), (ii) ReLU, (iii) Leaky ReLU, (iv) Clipped ReLU и (v) SatReLU. За моделите с Leaky ReLU, Clipped ReLU и SatReLU са разгледани по десет варианта на параметъра *const*. Съответно за Leaky ReLU това са стойностите от 0.01 до 0.1 със стъпка 0.01, а за Clipped ReLU и SatReLU от 10 до 100 със стъпка 10. За всеки от петте невронни модела и съответните стойности на параметъра *const* са извършени по десет независими обучения на рекурентния слой. За проверка поведението на различните модели на неврони в рекурентния слой е използвана DEAP базата данни. Използваните данни са от десет участника, като класификацията се извършва въз основа на изчислената нормализирана кратковременна енергия от 32 канален ЕЕГ сигнал. Целта на изследването е разпознаването между два класа: (а) клас *негативни* емоционални състояния и (б) обобщен клас *други*, включващ неутрални и позитивни емоционални състояния. Използваните данни са разделени в три групи: обучаващи данни – 20%; развойни данни – 20%; тестови данни – 60%.

5.4.2. Резултати



Фигура 5.4. Обобщени резултати от проведеното експериментално изследване с пет различни невронни модела на ЛРВНМ (точност %).

При подходящ избор на стойност за параметъра *const*, невронният модел с функцията SatReLU превъзхожда моделите с ReLU и Clipped ReLU и е конкурентен на моделите с активиращи функции Leaky ReLU и Sigmoid.

5.5. Обобщения и изводи

- Създаден и изследван е паралелен апаратен модел с висока разрядност на предавателната функция на ЛРВНМ, постигащ по-добри резултати от тези на класификатор с опорни вектори (SVM) [П.8].
- Създаденият апаратен модел показва резултати аналогични с програмната реализация на ЛРВНМ [П.8].
- Показана е необходимостта от компромис между ресурси, точност и време на изчисление [П.8].
- Създадена е нова разновидност на ЛРВНМ, използваща в локално рекурентния слой неврони с активираща функция от семейството на ReLU [П.7], [П.9].
- Подобрена е точността на класификация спрямо вече съществуващ модел на ЛРВНМ. Активираща функция Leaky ReLU показва най-добри резултати [П.7], [П.9].
- Създадена е нова активираща функция от семейството на ReLU, наречена SatReLU, която комбинира положителните характеристики на Leaky ReLU и Clipped ReLU. Активиращата функция SatReLU демонстрира конкурентни резултати с тези на Leaky ReLU [П.9].

5.6. Приноси към Глава 5

- Създадена е нова разновидност на локално рекурентната вероятностна невронна мрежа, която в локално рекурентния слой използва неврони с активираща функция от семейството на ReLU.

- Създадена е нова активираща функция от семейството на ReLU, наречена SatReLU, която комбинира положителните характеристики на Leaky ReLU и Clipped ReLU.
- Създаден е паралелен апаратен модел на локално рекурентната вероятностна невронна мрежа, с висока разрядност на предавателната функция в рекурентния слой.
- Валидирана е апаратна реализация на локално рекурентната вероятностна невронна мрежа, чрез сравнително експериментално изследване с (а) класификатор с опорни вектори (SVM) и (б) програмна реализация на локално рекурентната вероятностна невронна мрежа в среда на MATLAB.

Общи изводи и заключение

В дисертационният труд е създаден и изследван паралелен апаратен модел на ВНМ за класификацията на бинарни изображения. Предложени са промени в архитектурата на ВНМ позволяваща оптимизиране на използваните ресурси на FPGA при реализацията на апаратен модел на ВНМ за бинарни изображения. Реализираните промени предлагат приемлив компромис между постигната точност, заета площ и изчислителна скорост.

Апаратната реализация на невронна мрежа, може значително да подобри енергийната консумация на дадено приложение, а занижаване на необходимите ресурси за реализацията му и цялостната себестойност. За целта е разработен и изследван адаптиран метод за апроксимация на експоненциалната активираща функция при апаратна реализация на ВНМ. Адаптираният метод позволява намаляването на необходимия брой ресурси при апроксимацията на експоненциалната функция, като същевременно запазва класификационните способности на мрежата. Това от своя страна позволява увеличаване на броя неврони в скритите слоеве и подобряване класификационната точност. Също така са изследвани и оценени алтернативи на експоненциалната активираща функция, подходящи при апаратна реализация на ВНМ. Предложени са две алтернативни активиращи функции постигащи сходни резултати с тези на експоненциалната активираща функция, но с по-добри характеристики за апаратна реализация.

Изследвани и оценени са успеваемостта и бързината на деветнадесет конфигурации за обучение на ВНМ. Извършен е анализ от гледна точка на удобство при апаратна реализация на същите, във FPGA. В следствие е избран метод за обучение на ВНМ удобен при апаратна реализация във FPGA, което дава възможност за редуциране на необходимите ресурси и време за обучение.

Създаден и изследван е паралелен апаратен модел с висока разрядност на предавателната функция в ЛРВНМ. Апаратната реализация на невронната мрежа е валидирана чрез сравнително експериментално изследване с класификатор с опорни вектори (SVM) и програмна реализация на ЛРВНМ в среда на MATLAB. Създаденият апаратен модел показва по-висока точност при класификация при сравнение с класификатор с опорни вектори. Апаратния модел и програмната реализация на ЛРВНМ постигат сходни резултати, но апаратния е значително по-бърз. С цел по-добра адаптация на ЛРВНМ за апаратна реализация е създадена разновидност, която в локално рекурентния слой използва неврони с активираща функция от семейството на ReLU. Новосъздадената архитектура е подходяща за апаратна реализация, като същевременно постига и по-добри резултати от съществуващия програмен модел на ЛРВНМ.

Представените в дисертационния труд оптимизации и създадените нови модели предоставят възможност за реализацията на алгоритми на невронни структури за автономни преносими устройство с ниска консумация за медицински нужди или приложения изискващи минимална човешка намеса.

Основните резултати от изследванията и експериментите в дисертационния труд са публикувани в доклади и статии в национални и международни конференции, и списания, три от които са реферирани в международната научна база данни „Scopus“.

Приноси по дисертационния труд

Научни приноси:

1. Създадена е нова разновидност на локално рекурентната вероятностна невронна мрежа, която в локално рекурентния слой използва неврони с активираща функция от семейството на ReLU (Глава 5).
2. Създадена е нова активираща функция от семейството на ReLU, наречена SatReLU, която комбинира положителните характеристики на Leaky ReLU и Clipped ReLU (Глава 5).

Научно-приложни приноси:

3. Създаден е паралелен апаратен модел на локално рекурентната вероятностна невронна мрежа, с висока разрядност на предавателната функция в рекурентния слой (Глава 5).
4. Валидирана е апаратна реализация на локално рекурентната вероятностна невронна мрежа, чрез сравнително експериментално изследване с (а) класификатор с опорни вектори (SVM) и (б) програмна реализация на локално рекурентната вероятностна невронна мрежа в среда на MATLAB (Глава 5).
5. Създадена е методика, позволяваща оптимизиране на използваните FPGA ресурси при реализацията на апаратен модел на вероятностната невронна мрежа за бинарни изображения (Глава 2).
6. Подобрен е метод за апроксимация на експоненциалната активираща функция при апаратна реализация на вероятностната невронна мрежа (Глава 3).

Приложни приноси:

7. Разработен е приложно-оптимизиран паралелен апаратен модел на вероятностната невронна мрежа подходящ за бърза класификация на бинарни изображения (Глава 2).
8. Предложени, изследвани и оценени са алтернативи на експоненциалната активираща функция, подходящи при апаратна реализация на вероятностната невронна мрежа (Глава 3).
9. Дефинирани са критерии за удобство при апаратна реализация с FPGA и е избран метод за обучение на вероятностната невронна мрежа (Глава 4).

Публикации по дисертационния труд

[П.1] **N. Dukov**, T. Ganchev, D. Kovachev (2014). FPGA Implementation of Probabilistic Neural Network with the Altera Cyclone Family. Proc. Of the international scientific conference UNITECH-2014, 21-22 November 2014, Gabrovo, Bulgaria vol.1, pp.309-314.

[П.2] **N. Dukov**, T. Ganchev, D. Kovachev (2014). Empirical Study of PNN Sensitivity to the Exponential Function Approximation in FPGA Implementations. Annual of TU-Varna – 2014, pp.53-58.

[П.3] **Н. Дуков**, Т. Ганчев, Д. Ковачев (2015). Изследване на архитектури на изкуствени невронни мрежи за класификация на неврологични разстройства и сърдечни заболявания. Сборник Иновации и Бизнес 2015 – Научен форум „Приложни технологии свързани със здравето“, 9-10 Октомври 2015, Варна, България, стр.148-151.

[П.4] **N. Dukov**, T. Ganchev, D. Kovachev, M. N. Vrahatis (2015). Population size tradeoffs in DE and PSO-based methods for PNN training. Proc. Of the international scientific conference UNITECH-2015, 20-21 November 2015, Gabrovo, Bulgaria vol.1, pp.441-446.

[П.5] **N. Dukov**, T. Ganchev, D. Kovachev (2015). Reduced-complexity FPGA-implementation of PNN for classification of binary images. Annual of TU-Varna – 2015, pp.59-63.

[П.6] **N. Dukov** (2016). Comparative evaluation of Probability Density Estimators for the Probabilistic Neural Network. Proc. XXV International Scientific Conference Electronics - ET2016, 12-14 September 2016, Sozopol, Bulgaria, pp.21-24.

[П.7] **N. Dukov**, T. Ganchev (2017). Empirical study on ReLU based neuron models of the LRPNN. Proceedings of the 2017 BDAA Workshop, 13-14 October 2017, Varna, Bulgaria, pp.24-27.

[П.8] **N. Dukov**, T. Ganchev, D. Kovachev (2018). FPGA implementation of the Locally Recurrent Probabilistic Neural Network. Advances in Intelligent Systems and Computing, 679, pp. 419-428.

[П.9] **Н. Дуков**, Т. Ганчев (2018). Нови невронни модели на ЛРВНМ базирани на ReLU активиращи функции. V-та научна конференция с международно участие "Компютърни Науки и Технологии", 28-29 Септември 2018, Варна, България, стр.135-142

[П.10] **N. Dukov**, T. Ganchev, M. N. Vrahatis (2018). Comparative evaluation of DE and PSO-based methods for PNN training. Expert Systems With Applications – в процес на преразглеждане (under revision)

Декларация за оригиналност

от Николай Тинков Дуков

Декларирам, че представената от мен дисертация за получаване на образователна и научна степен „доктор“ на тема „Изследване на възможности за създаване на имитационни базови невронни структури и техните реализации“ е оригинален труд и съдържа оригинални резултати, получени при проведени от мен научни изследвания (с подкрепата и/или съдействието на научните ми ръководители).

Настоящата дисертация не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.

Дата:.....г.

Декларатор:.....

/Николай Т. Дуков/

Abstract

PhD Thesis Title: Study on the design and implementation of biologically motivated neural structures

submitted in partial fulfilment of the requirements for the degree Doctor of Philosophy

By: Nikolay Tinkov Dukov

The dissertation examines the possibilities of design and implementation of neural structures. Chapter 1 focuses on the specific properties and the challenges concerned with implementing neural networks. Moreover, it analyses the possibilities of implementing them in hardware. Furthermore, the technology for implementation is chosen and discussed. In Chapter 2 an adapted parallel hardware model of the probabilistic neural network (PNN) for the classification of binary images is presented. Additionally, steps for optimizing the design are offered. Chapter 3 deals with the analyses of a presented adapted method for approximation of the exponential activation function in hardware for the PNN. Moreover, alternative activation functions for PNN better suited for hardware implementation are explored and studied. Chapter 4 presents a study on nineteen algorithms for training the PNN. The algorithms are examined from hardware implementation point of view and a suitable algorithm is chosen. Finally, Chapter 5 focusses on a proposed hardware implementation of the locally recurrent probabilistic neural network (LRPNN) with high precision of the transfer function in the recurrent layer. Furthermore, a new neural architecture for the recurrent layer with ReLU functions is proposed as well as a new ReLU based activation function.