Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Τεχνικές Μηχανικής Μάθησης για το Διαδίκτυο των Πραγμάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΑΚΕΛΛΑΡΙΟΣ Γ. ΨΑΡΟΜΠΑΣ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης

Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης

Ε.ΔΙ.Π. Ε.Μ.Π.

Αθήνα, Ιούλιος 2019

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Τεχνικές Μηχανικής Μάθησης για το Διαδίκτυο των Πραγμάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΑΚΕΛΛΑΡΙΟΣ Γ. ΨΑΡΟΜΠΑΣ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Ε.ΔΙ.Π. Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Ιουλίου 2019.

Ανδρέας-Γεώργιος Σταφυλοπάτης	Παναγιώτης Τσανάκας	Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2019

Σακελλάριος Γ. Ψαρομπάς

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η Μηχανική Μάθηση αποτελεί τον τομέα της επιστήμης των υπολογιστών που αποσκοπεί στην μελέτη και κατασκευή συστημάτων τα οποία μέσω αυτοματοποιημένης ανάλυσης μετατρέπουν τα δεδομένα με τα οποία τροφοδοτούνται σε χρήσιμη πληροφορία. Το Διαδίκτυο των Πραγμάτων αποτελεί τη διασύνδεση συσκευών και πραγμάτων σε ένα κοινό δίκτυο με σκοπό την ευκολότερη ανταλλαγή δεδομένων μεταξύ τους αλλά και την εκμετάλλευση αυτής της επικοινωνίας για την εξαγωγή γνώσης χρήσιμης για τον άνθρωπο. Σκοπός της παρούσας εργασίας είναι η παρουσίαση μέρους του ερευνητικού έργου στο οποίο συνυπάρχουν οι παραπάνω κλάδοι. Η καταγραφή περιλαμβάνει το πεδίο εφαρμογής, τις μεθόδους και τους αλγορίθμους που χρησιμοποιήθηκαν για την επίλυση του εκάστοτε προβλήματος καθώς και την αξιολόγηση των αποτελεσμάτων τους.

Λέξεις κλειδιά

Διαδίκτυο των Πραγμάτων, Μηχανική Μάθηση, έξυπνο σπίτι, έξυπνη πόλη

Abstract

Machine Learning is the field of computer science that aims at designing and constructing systems which, through automated analysis, convert the data they provide into useful information.The Internet of Things is the interconnection of devices and things in a common network in order to facilitate the exchange of data between them, but also the exploitation of this communication to extract useful knowledge. The purpose of this thesis is to present those research areas in which the above disciplines coexist. The scope, methods and algorithms used to solve the problem, along with the evaluation the respective metrics, are being documented.

Key words

Internet of Things, Machine Learning, smart home, smart city

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του προπτυχιακού προγράμματος σπουδών της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου και σηματοδοτεί την ολοκλήρωση των σπουδών μου ενώ συγχρόνως αποτελεί το ερέθισμα για περαιτέρω έρευνα στο συγκεκριμένο αντικείμενο.

Προτού όμως αναφερθώ στη περιγραφή της εργασίας , θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς όλους αυτούς τους ανθρώπους των οποίων η συνεργασία και βοήθεια συνέβαλε καθοριστικά στην ολοκλήρωσή της.

Αρχικά θα ήθελα να απευθύνω τις ευχαριστίες μου στον επιβλέποντα κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, Καθηγητή Ε.Μ.Π, ο οποίος μου προσέφερε τη δυνατότητα να εκπονήσω την διπλωματική μου σε ένα αντικείμενο ιδιαίτερα ελκυστικό και ενδιαφέρον για μένα και να διευρύνω τις επιστημονικές μου γνώσεις. Παράλληλα θα ήθελα να ευχαριστήσω τους κ.κ. Παναγιώτη Τσανάκα, Καθηγητή Ε.Μ.Π και Γεώργιο Στάμου, Αναπληρωτή Καθηγητή Ε.Μ.Π για την τιμή που μου έκαναν να είναι μέλη της επιτροπής εξέτασης της διπλωματικής εργασίας.

Επίσης οφείλω να ευχαριστήσω τον κ. Γεώργιο Αλεξανδρίδη, Ε.ΔΙ.Π Ε.Μ.Π. για τη βοήθεια που προσέφερε κατά τη διάρκεια συγγραφής της συγκεκριμένης εργασίας. Οι συμβουλές και η καθοδήγηση του σε όλη τη διάρκεια της πορείας αυτής συνέβαλαν τα μέγιστα στην επίτευξη ενός πολύ σημαντικού για εμένα στόχου. Η εμπειρία, οι γνώσεις του καθώς και η προθυμία του να τις μοιραστεί μαζί μου, αποτέλεσαν όχι μόνο καθοριστικό παράγοντα ολοκλήρωσης της παρούσας εργασίας αλλά και ερέθισμα για περαιτέρω προβληματισμό και έρευνα.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, η στήριξη της οποία ήταν καταλυτικός παράγοντας της ολοκλήρωσης των σπουδών μου στο Εθνικό Μετσόβιο Πολυτεχνείο.

Σακελλάριος Γ. Ψαρομπάς,

Αθήνα, 25η Ιουλίου 2019

Περιεχόμενα

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Διαδίκτυο των Πραγμάτων

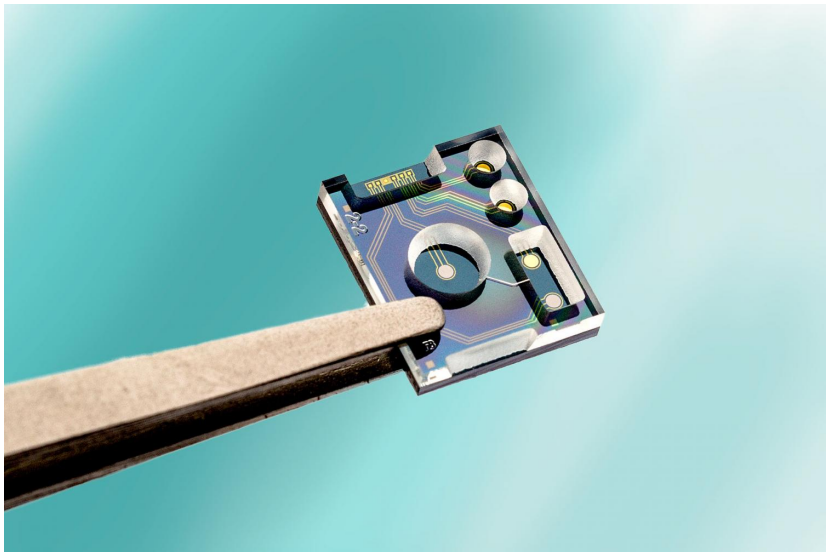
Η ραγδαία εξέλιξη της τεχνολογίας τα τελευταία χρόνια έχει καταστήσει την επικοινωνία μεταξύ των συσκευών πιο εύκολη από ποτέ. Το *Διαδίκτυο των Πραγμάτων* (Internet of Things ή IoT) αποτελεί τη διασύνδεση συσκευών και πραγμάτων σε ένα κοινό δίκτυο με σκοπό την ευκολότερη ανταλλαγή δεδομένων μεταξύ τους αλλά και την εκμετάλλευση αυτής της επικοινωνίας για την εξαγωγή γνώσης χρήσιμης για τον άνθρωπο. Ο όρος αντικείμενο μπορεί να αναφέρεται σε αισθητήρες, κινητά τηλέφωνα αλλά και γενικότερα οτιδήποτε έχει τη δυνατότητα σύνδεσης στο Διαδίκτυο. Θα μπορούσαμε να πούμε ότι ο τελικός στόχος είναι η ενσωμάτωση-απεικόνιση του πραγματικού κόσμου στον ψηφιακό.

Οι προκλήσεις που προκύπτουν για την πραγματοποίηση των παραπάνω είναι πολλές και σε ποικίλα επίπεδα. Σε αυτό της αρχιτεκτονικής υπάρχει η ανάγκη για δυνατότητα σύνδεσης πολλών συσκευών με τρόπο που θα μειώνει όσο είναι δυνατόν την κίνηση στο δίκτυο.

Ο τομέας των δεδομένων συγκεντρώνει το περισσότερο ενδιαφέρον για την παρούσα εργασία. Οι συσκευές στο δίκτυο παράγουν μεγάλο όγκο δεδομένων τα οποία πρέπει να αποθηκευτούν και να επεξεργαστούν. Το γεγονός ότι υπάρχει πλήθος διαφορετικών πηγών εισάγει μια ετερογένεια σε αυτά η οποία δυσχεραίνει τη διαδικασία. Η *Μηχανική Μάθηση* (Machine Learning ή ML) είναι ένα πεδίο του οποίου οι τεχνικές χρησιμοποιούνται για την εξαγωγή χρήσιμων συμπερασμάτων και γνώσης από τα υπάρχοντα δεδομένα. Στα επόμενα κεφάλαια θα παρουσιάσουμε αλγορίθμους που έχουν χρησιμοποιηθεί για συγκεκριμένες διεργασίες.

Επίσης υπάρχουν ακόμα ζητήματα όπως αυτό της ασφάλειας και των προσωπικών δεδομένων. Πολλές από τις πληροφορίες που κυκλοφορούν στο δίκτυο είναι «ευαίσθητες» και πρέπει να προφυλαχθούν. Το μεγάλο πλήθος, αλλά και η μικρή υπολογιστική δύναμη των συσκευών αποτελούν τροχοπέδη προς την κατεύθυνση αυτή, αφήνοντας παράλληλα και ανοιχτό το πεδίο της έρευνας.

Το IoT βρίσκει εφαρμογή σε μια πληθώρα από κλάδους που εκμεταλλεύονται σε πρακτικό επίπεδο τα όσα αυτό προσφέρει. Πιο χαρακτηριστικό και πολυεπίπεδο παράδειγμα είναι αυτό της «έξυπνης πόλης». Τα δεδομένα που συλλέγονται μέσα από αισθητήρες (Σχήμα



Σχήμα 1.1: Αισθητήρας

1.1) χρησιμοποιούνται για τη διαχείριση της κυκλοφοριακής κίνησης, τον καλύτερο έλεγχο της ενέργειας που καταναλώνεται αλλά και τη γενικότερη παρακολούθηση του περιβάλλοντος με συστηματικό τρόπο, τόσο για θέματα όπως αυτό της μόλυνσης αλλά και την πρόληψη κινδύνων.

Τα «έξυπνα κτίρια» επίσης μέσω αισθητήρων προσπαθούν να βελτιστοποιήσουν παράγοντες που αφορούν την λειτουργία τους αλλά και την καλύτερη εξυπηρέτηση όσων δραστηριοποιούνται σε αυτά. Η πρόβλεψη των ενεργειακών αναγκών σε συνάρτηση με τις καιρικές συνθήκες είναι ένα τέτοιο παράδειγμα.

Ένας ακόμα τομέας με μεγάλο ενδιαφέρον για το IoT είναι αυτός της υγείας. Η δυνατότητα για απομακρυσμένο και έγκαιρο έλεγχο των ασθενών αποτελεί σημαντικό παράγοντα που μπορεί να βοηθήσει στην ποιότητα ζωής. Όπως θα δούμε και στη συνέχεια, ο συγκεκριμένος κλάδος προσελκύει σημαντικό κομμάτι του ερευνητικού ενδιαφέροντος.

1.2 Μηχανική Μάθηση

Η μηχανική μάθηση είναι το πεδίο της επιστήμης των υπολογιστών που ασχολείται με την δημιουργία μεθόδων ανάλυσης, με τις οποίες ένας υπολογιστής αποσκοπεί στην απόκτηση γνώσης από ένα σύνολο δεδομένων. Χρησιμοποιείται ως εργαλείο από πολλούς κλάδους επαγγελματιών όπως μηχανικοί, στατιστικολόγοι και οικονομολόγοι, αφού μπορεί να εξάγει χρήσιμα συμπεράσματα και προβλέψεις ανεξάρτητα από τον τύπο των δεδομένων με τα οποία τροφοδοτείται.

Δεδομένου ενός συστήματος και ενός *συνόλου δεδομένων* (dataset), η μηχανική μάθηση, με τη χρήση διαφόρων αλγορίθμων, αρκετοί από τους οποίους θα εξεταστούν στα επόμενα Κεφάλαια, παράγει ένα μοντέλο. Σκοπός είναι για μελλοντικές εισόδους του συστήματος να

μπορούν να προβλεφθούν οι αντίστοιχες *ετικέτες* (labels) τους, εκμεταλλευόμενοι τη γνώση που βρίσκεται μέσα στα δεδομένα. Για τη δημιουργία του μοντέλου, το σύνολο δεδομένων χωρίζεται σε δύο μέρη; *εκπαίδευσης* (train) και *ελέγχου* (test), με το πρώτο να χρησιμοποιείται για την εκπαίδευση και το δεύτερο για την αποτίμηση του παραγόμενου μοντέλου.

Με βάση τα δεδομένα εισόδου γίνεται διαχωρισμός σε δύο βασικές κατηγορίες. Στην *επιβλεπόμενη μάθηση* (supervised learning), το *σύνολο εκπαίδευσης* (training set) περιέχει τις επιθυμητές ετικέτες για κάθε είσοδό του. Αντίθετα, στη *μη-επιβλεπόμενη μάθηση* (unsupervised learning) οι είσοδοι δεν συνοδεύονται από αντίστοιχες ετικέτες, ενώ είναι πιθανό κάποια προβλήματα να αποτελούν συνδυασμό των παραπάνω περιπτώσεων. Επίσης υπάρχει και η *ανατροφοδοτούμενη μάθηση* (reinforcement learning), στην οποία τα δεδομένα έρχονται σε πραγματικό χρόνο με τη μορφή ανατροφοδότησης.

Διαχωρισμός μπορεί να γίνει και με βάση τον τύπο των ετικετών, ανάλογα με το αν αυτές είναι ποιοτικές ή ποσοτικές. Στην πρώτη περίπτωση αναφερόμαστε σε μη διατεταγμένες μεταβλητές που μπορούν να κατηγοριοποιηθούν σε κλάσεις ενώ στη δεύτερη σε τιμές σε ένα συνεχές φάσμα διατεταγμένων μεταβλητών. Τα μοντέλα που προβλέπουν ποσοτικές μεταβλητές ονομάζονται *παλινδρόμησης* (regression) ενώ αυτά που προβλέπουν ποιοτικές ονομάζονται *ταξινόμησης* (classification). Μια ακόμη κατηγορία είναι η *συστηματοποίηση* (clustering), η οποία μοιάζει με την ταξινόμηση με τη διαφορά όμως ότι οι υπάρχουσες κλάσεις δεν είναι εκ των προτέρων γνωστές και έτσι ό αλγόριθμος θα πρέπει να κάνει από μόνος του την ομαδοποίηση.

1.3 Δομή της εργασίας

Η εργασία στα επόμενα κεφάλαια ακολουθεί την εξής δομή. Στο Κεφάλαιο 2 πραγματοποιείται μια αναλυτικότερη παρουσίαση του IoT. Κατά τη διαδικασία αυτή εμβαθύνουμε στα πεδία που βρίσκει εφαρμογή ενώ παράλληλα παραθέτουμε και αναφορές από τις αντίστοιχες υλοποιήσεις και ερευνητικό έργο. Στο Κεφάλαιο 3 ακολουθούμε την ίδια διαδικασία και για την μηχανική μάθηση. Αφού ομαδοποιήσουμε της εφαρμογές με βάση το έργο που επιλύουν, παραθέτουμε τους αλγορίθμους που χρησιμοποιούνται για το σκοπό αυτό. Στο Κεφάλαιο 4 παρουσιάζουμε τα δεδομένα που έχουν χρησιμοποιηθεί, καθώς και την μορφή που έχουν αυτά. Επίσης ασχολούμαστε με τις μετρικές απόδοσης που έχουν χρησιμοποιηθεί στις εργασίες που ερευνήσαμε καθώς και με μία θεωρητική παρουσίαση αυτών. Τέλος, στο Κεφάλαιο 5 παραθέτουμε τα συμπεράσματα που προκύπτουν από το σύνολο της εργασίας καθώς και τις μελλοντικές κατευθύνσεις που ενδέχεται να ακολουθήσει το πεδίο του IoT.

Κεφάλαιο 2

Διαδίκτυο των Πραγμάτων

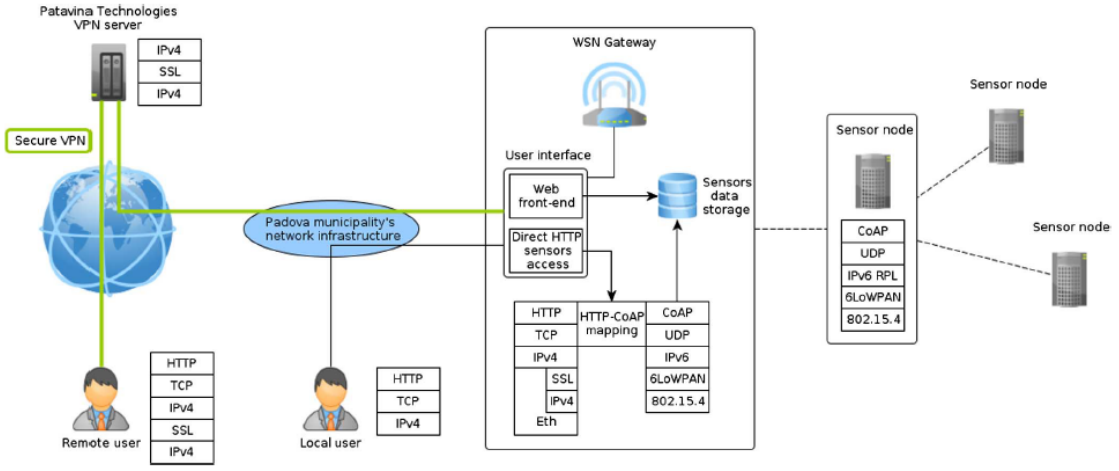
Σε αυτό το Κεφάλαιο θα επιχειρήσουμε μια αναλυτικότερη παρουσίαση του IoT. Κάθε υποενότητα θα ασχολείται με ένα πεδίο στο οποίο αυτό βρίσκει πρακτική εφαρμογή. Αφού αναφερθούν τα γενικά στοιχεία του πεδίου και οι στόχοι του, θα παρουσιαστούν και σχετικές ερευνητικές εργασίες.

2.1 Έξυπνη Πόλη

Ο όρος *«έξυπνος»* (smart) χρησιμοποιείται κατά κόρον σε τεχνολογικά ζητήματα τη σημερινή εποχή, χωρίς όμως να είναι πλήρως ξεκάθαρη η ουσιαστική έννοια του. Στην περίπτωση της *«έξυπνης πόλης»* (smart city) έχει να κάνει με την αυτόματη συλλογή και ανάλυση δεδομένων που αφορούν την πόλη. Τα δεδομένα αυτά προέρχονται κυρίως από αισθητήρες αλλά και άλλες πηγές (για παράδειγμα κινητά τηλέφωνα) που βοηθούν στην αντιμετώπιση σοβαρών θεμάτων που προκύπτουν στην καθημερινότητα. Με αυτό τον τρόπο οι ιθύνοντες έχουν τη δυνατότητα να βελτιώσουν την ποιότητα ζωής των πολιτών αλλά και τις παρεχόμενες υπηρεσίες προς αυτούς. Πιο συγκεκριμένα, εφαρμογές βρίσκουμε σε τομείς όπως η διαχείριση της ηλεκτρικής ενέργειας που καταναλώνεται σε μια πόλη. Με την βοήθεια των κατάλληλων δεδομένων αυτή μπορεί να προβλεφθεί αλλά και να περιορισθεί προς όφελος της οικονομίας, αλλά και του περιβάλλοντος. Μεγάλη επίδραση μπορεί να υπάρξει και στους δρόμους της πόλης με καλύτερο έλεγχο της κυκλοφορίας, της στάθμευσης αλλά και της αναβάθμισης των παρεχομένων υπηρεσιών στα μέσα μαζικής μεταφοράς. Επίσης βοήθεια μπορεί να υπάρξει σε θέματα δημόσιας ασφάλειας, στη λειτουργία δημοσίων υπηρεσιών και νοσοκομείων, στη διαχείριση των απορριμάτων αλλά και στον μελλοντικό προγραμματισμό για την εξέλιξη της πόλης.

Για την επίτευξη πρακτικής υλοποίησης των παραπάνω ιδεών, υπάρχουν μια σειρά από προκλήσεις που πρέπει να αντιμετωπιστούν. Αυτές έχουν να κάνουν κατά κύριο λόγο με την ετερογένεια των δεδομένων αλλά και με την επικοινωνία των συσκευών, οι οποίες συνήθως είναι περιορισμένων υπολογιστικών δυνατοτήτων. Για την αντιμετώπισή τους στο [Zane12] προτείνεται η υλοποίηση μιας αρχιτεκτονικής (Σχήμα 2.1) βασισμένης στο RESTful API.

Η ιδέα αυτή εφαρμόζεται στην πόλη της Πάντοβα, όπου εγκαθίστανται αισθητήρες για τον έλεγχο της φωτεινότητας. Μάλιστα οι αισθητήρες αυτοί καταγράφουν και άλλα στοιχεία σχετικά με τη θερμοκρασία, την υγρασία αλλά και την ατμοσφαιρική ρύπανση, ο συνδυασμός των οποίων επιτρέπει την εξαγωγή χρήσιμων συμπερασμάτων σχετικά με τα δρώμενα της πόλης.



Σχήμα 2.1: Αρχιτεκτονική βασισμένη στο RESTful API [Zane12]

Μία από τις πλέον σημαντικές και πρωτοπόρες μελέτες στα πλαίσια της «έξυπνης πόλης» είναι αυτή του SmartSantander [Sanc14]. Η σπουδαιότητα σε τέτοια εγχειρήματα έγκειται στο γεγονός ότι δεν εστιάζουν σε μία μόνο εφαρμογή, αλλά εξελίσσουν συνολικά το IoT. Το πεδίο της πόλης αποτελεί ένα ζωντανό περιβάλλον και συνεπώς προσφέρεται για πιο αξιόπιστα πειράματα σε σχέση με ένα εργαστήριο. Εκεί μπορούν να δοκιμαστούν υπηρεσίες, αλλά ακόμα και αρχιτεκτονικές για την υποδομή, κατευθείαν σε πραγματικές συνθήκες μεγάλης κλίμακας.

Ειδικά το τελευταίο αποτελεί από τους σημαντικότερους παράγοντες τόσο σε ερευνητικό όσο και πρακτικό επίπεδο. Η πόλη του Σανταντέρ παρέχει τη δυνατότητα να διαχειριστούμε κόμβους της τάξης των μερικών χιλιάδων, τη στιγμή που σε περιβάλλον εργαστηρίου αυτός ο αριθμός θα έπεφτε σε μερικές εκατοντάδες. Με δεδομένο ότι οι απαιτήσεις στον πραγματικό κόσμο ισούνται ή και ξεπερνούν αυτές της πρώτης περίπτωσης, γίνεται εύκολα αντιληπτή η χρησιμότητα να μπορούμε να εργαστούμε και να πειραματιστούμε σε ένα τέτοιο πεδίο.

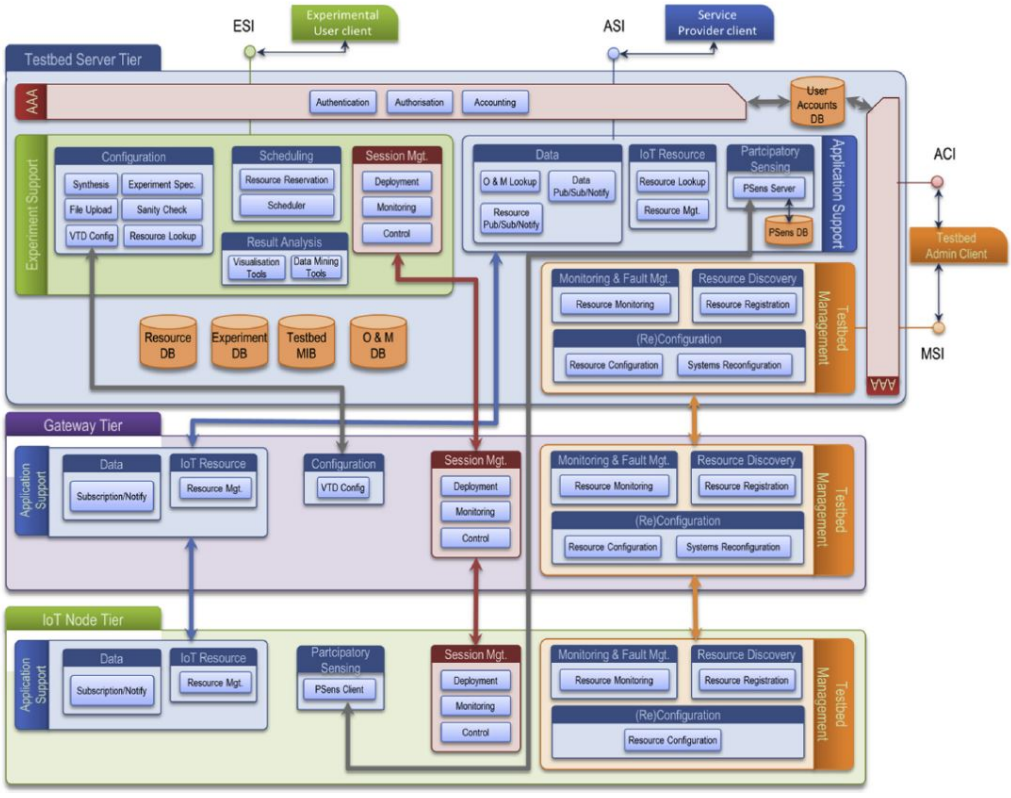
Ένα ακόμα πλεονέκτημα σε σχέση με το εργαστήριο είναι η φυσική ύπαρξη της έννοιας του χώρου και η δυνατότητα κίνησης σε αυτόν. Οι αισθητήρες μπορούν να αποκτούν κίνηση στο χώρο προσαρτημένοι πάνω σε οχήματα όπως τα μέσα μαζικής μεταφοράς και να καταγράφουν με φυσικό τρόπο και όχι μέσω διαδικασίας προσομοίωσης.

Επίσης το σύστημα δεν έχει ως χρήστες μόνο τους ερευνητές. Σε αυτό πρόσβαση έχουν και δημιουργοί οι οποίοι μπορούν να σχεδιάσουν εφαρμογές για εμπορική χρήση στα πλαίσια

της «έξυπνης πόλης». Ως χρήστες συμμετέχουν όμως και οι πολίτες, γεγονός που βοηθάει στην αξιολόγηση και των εντοπισμό προβλημάτων.

Η αρχιτεκτονική που χρησιμοποιείται αποτελείται από τρία επίπεδα όπως φαίνεται και στο Σχήμα 2.2. Στο επίπεδο κόμβου, που είναι και το χαμηλότερο, οι συσκευές είναι περιορισμένων δυνατοτήτων και λειτουργούν είτε ως αισθητήρες είτε ως ενεργοποιητές. Κατά το σχεδιασμό τους και την τοποθέτηση τους πρέπει να ληφθούν υπόψιν οι ενεργειακές τους απαιτήσεις καθώς και οι περιβαλλοντικές συνθήκες στις οποίες θα εκτεθούν. Το δεύτερο επίπεδο είναι υπεύθυνο για τη σύνδεση των συσκευών σε υποδομή δικτύου. Τέλος, το επίπεδο διακομιστή αποτελείται από πιο ισχυρά μηχανήματα που λειτουργούν ως αποθηκευτικό μέσο των δεδομένων αλλά και ως εξυπηρετητής για της εφαρμογές. Επίσης ο διακομιστής είναι υπεύθυνος για την διαθεσιμότητα και την αξιοπιστία του δικτύου.

Η αρχιτεκτονική δεν επηρεάζεται από τον τύπο και τον τρόπο σύνδεσης των κόμβων. Η ετερογένεια σε επίπεδο τόσο υλικού όσο και δεδομένων είναι από τα στοιχεία που καλείται να αντιμετωπίσει ένα σύστημα τέτοιου σκοπού. Επίσης σκοπός των σχεδιαστών είναι η λειτουργία να είναι εφικτή με όσο το δυνατόν μικρότερη ανθρώπινη παρέμβαση.



Σχήμα 2.2: Αρχιτεκτονική SmartSantander [Sanc14]

Το SmartSantander αξιοποιεί τα παραπάνω σε μια σειρά πρακτικών εφαρμογών. Πρώτη από αυτές είναι η καταγραφή και η πρόβλεψη περιβαλλοντικών δεικτών μέσω των αισθητήρων που διαθέτει το σύστημα καθώς και η αυτόματη ρύθμιση της άρδευσης σε πάρκα και σε

κήπους. Η χρήση *επαυξημένης πραγματικότητας* (augmented reality) κάνει την πόλη πιο προσβάσιμη και φιλική στον επισκέπτη ενώ προς την κατεύθυνση αυτή βοηθάει και το σύστημα υποστήριξης οδήγησης και στάθμευσης. Επίσης οι πολίτες μπορούν να βοηθούν στη συλλογή των δεδομένων μέσω των κινητών τους τηλεφώνων και άλλων συσκευών αλλά και να καταγράφουν συγκεκριμένα γεγονότα τα οποία είναι ορατά από τους υπόλοιπους χρήστες.

Στο [Hrom15] πραγματοποείται ανάλυση της ατμοσφαιρικής ρύπανσης για την πόλη του Ζάγκρεμπ. Η συλλογή των δεδομένων γίνεται με τη βοήθεια πολιτών οι οποίοι φέρουν αισθητήρες στα ρούχα ή τα οχήματά τους. Οι αισθητήρες, οι οποίοι έχουν πολύ φθηνό κόστος παραγωγής επικοινωνούν με τα κινητά τηλέφωνα των χρηστών τα οποία αποστέλλουν τα δεδομένα στον κεντρικό διακομιστή. Με αυτό τον τρόπο γίνεται μια χωρική και χρονική καταγραφή των επιπέδων των ρύπων στην ατμόσφαιρα. Απώτερος σκοπός των ερευνητών είναι ο εντοπισμός συσχέτισης μεταξύ των μεταβλητών που επηρεάζουν την ρύπανση.

Στο [Jin14] παρουσιάζεται ένα σύστημα για την παρακολούθηση της ηχορύπανσης στην πόλη της Μελβούρνης. Στο σύστημα εκτός από τους σταθερούς αισθητήρες, με δεδομένα συνεισφέρουν και οι πολίτες μέσω αισθητήρων στα κινητά τηλέφωνα και τα οχήματά τους. Με αυτό τον τρόπο συμπληρώνονται τα χωρικά κενά στα δεδομένα παρέχοντας έτσι μια πιο ολοκληρωμένη εικόνα. Η αρχιτεκτονική προσεγγίζει υλοποίηση με 3 επίπεδα αισθητήρων και αναμεταδοτών με σκοπό την αποτελεσματικότερη μετάδοση των δεδομένων.

Ένα από τα βασικά χαρακτηριστικά όλων των παραπάνω εφαρμογών είναι ο μεγάλος όγκος των δεδομένων προς επεξεργασία. Στο [Rath16] προτείνεται μια αρχιτεκτονική τεσσάρων επιπέδων με τη χρήση τεχνολογιών *Μεγάλων Δεδομένων* (Big Data), όπως το *Apache Hadoop* [Vavi13] και το *Apache Spark* [Zaha16] για την αποθήκευση και ανάλυση των δεδομένων.

2.2 Έξυπνο Σπίτι

Ο όρος *«έξυπνο σπίτι»* (smart home) μπορεί να αναφέρεται τόσο σε σπίτια όσο όμως και γενικότερα σε κτίρια. Με την χρήση των IoT τεχνολογιών αποσκοπεί στην βελτίωση της διαβίωσης ή παραμονής των ανθρώπων μέσα σε αυτό. Οι αισθητήρες και οι τεχνικές μηχανικής μάθησης μπορούν να υλοποιήσουν μια σειρά από αυτοματοποιημένες λειτουργίες που διευκολύνουν τον άνθρωπο, αυξάνουν τα επίπεδα ασφαλείας και καθιστούν το κτίριο πιο φιλικό στο περιβάλλον.

Στο [Derg14] προτείνεται μια υλοποίηση για την πρόβλεψη κατανάλωσης ηλεκτρικού ρεύματος ενός κτιρίου. Για το σκοπό αυτό χρησιμοποιούνται ελεύθερα προσβάσιμες πηγές από το Διαδίκτυο σχετικές με τις καιρικές συνθήκες καθώς και αισθητήρες οι οποί ελέγχουν την τρέχουσα κατανάλωση. Επίσης λόγω του γεγονότος ότι τα δεδομένα προέρχονται από πολλές διαφορετικές πηγές, το σύστημα είναι υπεύθυνο για την κατάλληλη μεταφορά

και αποθήκευσή τους, καθώς και για την επιλογή της πηγής που θα δώσει τις πιο ακριβείς προβλέψεις.

Η κατανάλωση ενέργειας αποτελεί το αντικείμενο και του [Jakk10]. Σε αντίθεση όμως με την προηγούμενη περίπτωση, δεν γίνεται προσπάθεια πρόβλεψης της κατανάλωσης αλλά να εντοπίζεται μη φυσιολογική κίνηση σε αυτή. Για να επιτευχθεί ο σκοπός αυτός χρησιμοποιούνται στατιστικές μεθοδολογίες αλλά και τεχνικές συσταδοποίησης. Με τον τρόπο αυτό μπορούν να προληφθούν έγκαιρα δυνητικοί κίνδυνοι που προέρχονται από αστοχίες ηλεκτρικών συσκευών αλλά και να αποτελέσει ένα μέτρο για την μείωση της συνολικής κατανάλωσης.

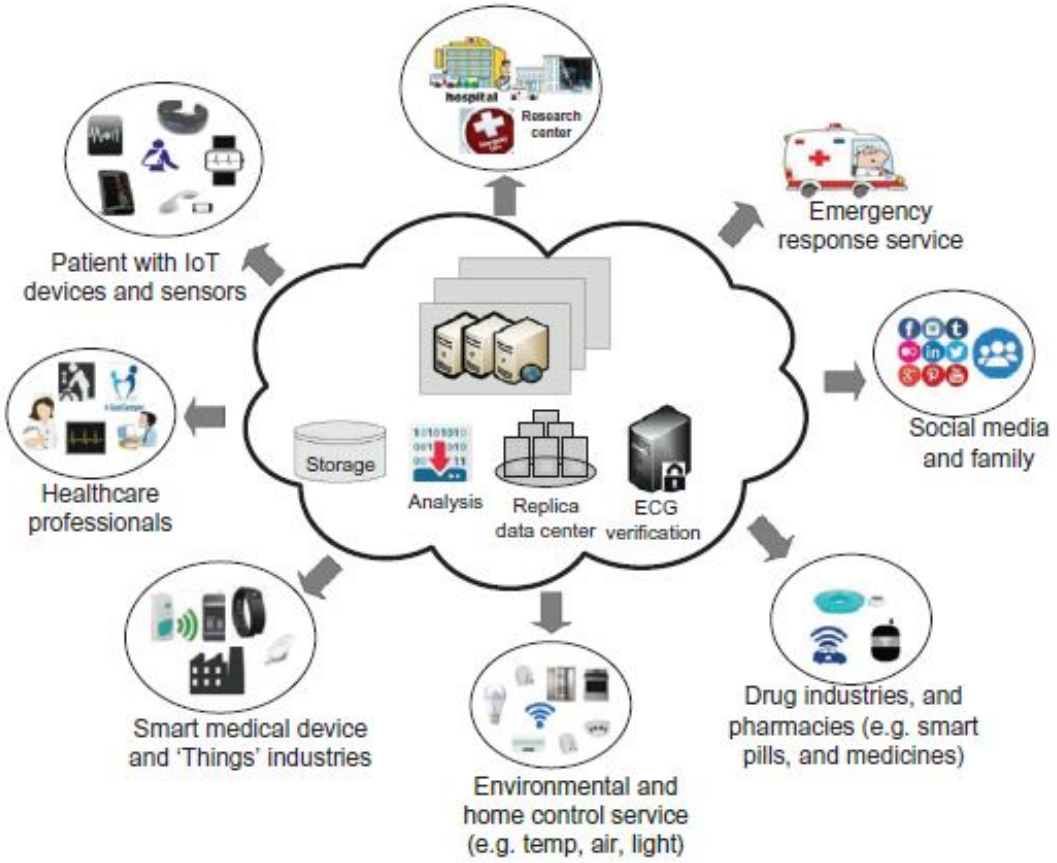
Ένα από τα προβλήματα που μπορούν να προκύψουν σε εφαρμογές όπως οι παραπάνω είναι ορισμένοι αισθητήρες, λόγω σφάλματος ή παρεμβολών από το περιβάλλον, να παρουσιάζουν λανθασμένες τιμές. Για αυτό το σκοπό στο [Mone13] παρουσιάζεται μία μέθοδος για τον έλεγχο της εγκυρότητας των αισθητήρων. Αν και υπάρχουν πολλοί τρόποι για να επιτευχθεί το εν λόγω αποτέλεσμα, στη συγκεκριμένη εργασία χρησιμοποιούνται στατιστικά μοντέλα. Οι αισθητήρες αντί να αντιμετωπίζονται μεμονωμένα, ελέγχονται σαν σύνολο και ερευνάται η σχέση που έχουν μεταξύ τους.

Στο [Feng17] προτείνεται η χρήση του *Cognitive Dynamic System* (CDS). Με αυτό τον τρόπο το ίδιο το σπίτι μπορεί να αντιλαμβάνεται τι συμβαίνει μέσα στο περιβάλλον του μέσω των αισθητήρων που έχουν εγκατασταθεί αλλά και να αντιδρά σε αυτά τα συμβάντα με τις κινήσεις που του έχουν καθοριστεί.

2.3 Υπηρεσίες Υγείας

Ο τομέας της υγείας είναι από αυτούς που μπορούν να επωφεληθούν ποικιλοτρόπως και σε μεγάλο βαθμό από το IoT και κατ' επέκταση προσελκύει και έντονο ερευνητικό ενδιαφέρον. Η δυνατότητα για παρακολούθηση σε πραγματικό χρόνο του ασθενούς μέσω αισθητήρων μπορεί να βελτιώσει σημαντικά τόσο το κομμάτι της πρόληψης όσο και το κομμάτι της αντιμετώπισης πληθώρας παθήσεων. Ακόμα και σε περιπτώσεις που δεν υπάρχει κάποια συγκεκριμένη πάθηση, ιατρικοί δείκτες όπως οι καρδιακοί παλμοί μπορούν να υποδείξουν την ανάγκη για αλλαγή στις συνήθειες ή την διατροφή του ατόμου. Επίσης η συστηματική καταγραφή βοηθάει τόσο στην στοιχειοθέτηση καλύτερου ιστορικού τόσο σε ατομικό επίπεδο για τον ασθενή όσο και σε συλλογικό που δύναται να χρησιμοποιηθεί για ερευνητικούς σκοπούς. Τέλος, μέσα σε ένα πλαίσιο όπου σημαντικός αριθμός ατόμων έχει τέτοιου είδους παρακολούθηση, είναι πιο εύκολο να εντοπισθούν και να αντιμετωπιστούν επιδημίες. Στο Σχήμα 2.3 παρουσιάζεται ένα γενικό πλαίσιο στο οποίο θα μπορούσε λειτουργεί μια εφαρμογή με επίκεντρο την υγεία στο IoT.

Στο [Pand17] μελετάται την εμφάνιση στρες στα άτομο μέσω της παρακολούθησης των



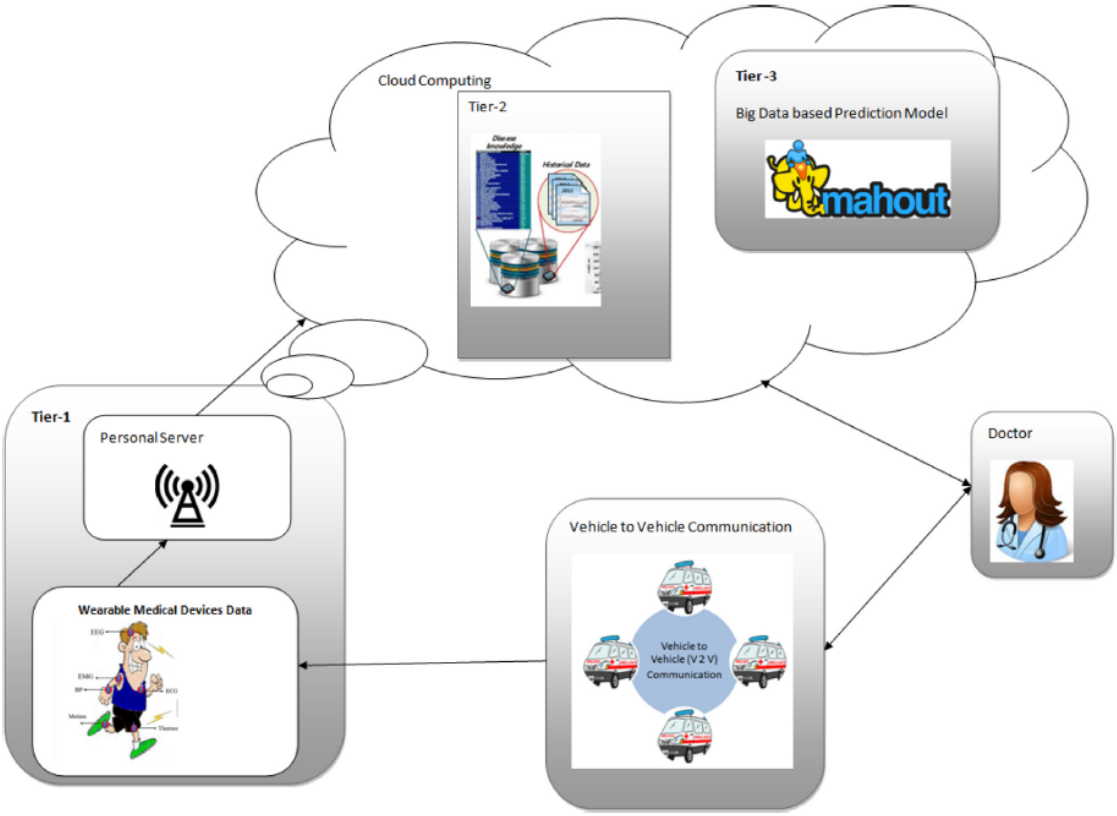
Σχήμα 2.3: Πλαίσιο λειτουργίας εφαρμογής για την υγεία [Hoss16]

καρδιακών τους παλμών. Το IoT προσφέρει την υποδομή ώστε να μπορεί να γίνει η μετάδοση των δεδομένων ενώ η Μηχανική Μάθηση παρέχει τις τεχνικές με τις οποίες γίνεται η ανάλυσή τους. Το σύστημα χρειάζεται μια αρχική ρύθμιση που είναι προσωπική για κάθε χρήστη του και μετά από αυτό είναι σε θέση να εντοπίζει πιθανούς κινδύνους για την υγεία του.

Στο [Hass15] παρουσιάζεται η αρχιτεκτονική ενός συστήματος που επιτελεί εργασία παρόμοια με το παραπάνω. Αισθητήρες που βρίσκονται πάνω στο σώμα και τα ρούχα του ατόμου καταγράφουν μία σειρά από βιομετρικούς δείκτες (π.χ θερμοκρασία σώματος). Τα δεδομένα αυτά συλλέγονται σε ένα ενδιάμεσο σταθμό το ρόλο του οποίου συνήθως παίζει το κινητό τηλέφωνο του χρήστη και από εκεί αποστέλλονται στον κεντρικό διακομιστή. Ο τελευταίος είναι υπεύθυνος για την αποθήκευσή τους, την ανάλυση τους αλλά και τη μετατροπή των τελικών αποτελεσμάτων σε μορφή εύκολα αναγνώσιμη από τον άνθρωπο. Εναλλακτικά μπορεί να υπάρχει και ένας ενδιάμεσος σταθμός, ο οποίος να έχει τη δυνατότητα εκτέλεσης ορισμένων υπολογιστικών διαδικασιών αυξάνοντας έτσι το χρόνο απόκρισης αλλά και μειώνοντας τη συμφόρηση στο δίκτυο.

Στο [Kuma18] προτείνεται ένα σύστημα αρκετά παρόμοιο με αυτό του [Hass15], αλλά

λαμβάνοντας υπόψιν μια ακόμα παράμετρο. Το μέγεθος των δεδομένων που παράγονται προς αποθήκευση και επεξεργασία είναι πολύ μεγάλο και το σύστημα στο σύνολο του πρέπει να είναι κλιμακώσιμο ως προς αυτά. Για το σκοπό αυτό χρησιμοποιείται η *Apache Hbase*[1] για την αποθήκευση τους με κατανεμημένο τρόπο. Επίσης η επεξεργασία γίνεται και αυτή σε κατανεμημένο περιβάλλον με χρήση της βιβλιοθήκης *Apache Mahout* [Lyub16]. Στο Σχήμα 2.4 φαίνεται η διαδρομή που ακολουθούν τα δεδομένα για αποθήκευση και επεξεργασία.



Σχήμα 2.4: Αρχιτεκτονική προτεινόμενου συστήματος [Kuma18]

Η αρχιτεκτονική που προτείνεται στο [Mano18] είναι και αυτή παρόμοια και χρησιμοποιεί υπηρεσίες των *Amazon Web Services* (AWS)[2] καθώς και την Apache Hbase για την κατανεμημένη αποθήκευση των δεδομένων των ασθενών. Η διαφοροποίηση της έγκειται στο γεγονός ότι εισάγει στην διαδικασία την έννοια του *fog computing* [Bono12], το οποίο σημαίνει ότι το σύνολο της επεξεργασίας δεν εκτελείται εξ' ολοκλήρου στον κεντρικό διακομιστή αλλά μέρος αυτής διαμοιράζεται και σε άλλα σημεία του δικτύου. Αυτό το γεγονός έχει σαν αποτέλεσμα να είναι μικρότερη η συμφόρηση στο δίκτυο και τα αποτελέσματα να λαμβάνονται πιο γρήγορα.

[1] `https://hbase.apache.org/`
[2] `https://aws.amazon.com/`

2.4 Διαχείριση κυκλοφοριακής κίνησης

Ένα από τα πλέον σημαντικά ζητήματα που έχουν να αντιμετωπίσουν τα αστικά κέντρα είναι η διαχείριση της κυκλοφορίας των οχημάτων μέσα σε αυτά. Μάλιστα για αυτό το λόγο έχει υπάρξει έντονο ερευνητικό ενδιαφέρον γύρω από αυτό και συνεπώς αξίζει να το εξετάσουμε ξεχωριστά από το γενικότερο πλαίσιο της «έξυπνης πόλης» στο οποίο εντάσσεται. Τα οφέλη που μπορούν να προκύψουν αφορούν τόσο θέματα τα οποία εξελίσσονται σε πραγματικό χρόνο όσο και μακροχρόνια. Η γνώση της κατάστασης του οδικού δικτύου μπορεί να κατευθύνει τους οδηγούς με τέτοιο τρόπο ώστε να αποφευχθεί η συμφόρηση που οφείλεται είτε σε αυξημένη κίνηση είτε σε εξωγενείς παράγοντες όπως ατυχήματα ή τεχνικά προβλήματα του δικτύου. Με αυτό τον τρόπο εξοικονομείται χρόνος, ενέργεια αλλά και μειώνονται οι περιβαλλοντικοί ρύποι. Η συγκέντρωση αυτών των στοιχείων σε βάθος χρόνου μπορεί να βοηθήσει και στον προγραμματισμό για την ανάπτυξη της πόλης, καθώς προσφέρουν μια ξεκάθαρη εικόνα σχετικά με τις ανάγκες για περαιτέρω οδικές υποδομές. Επίσης ένας ακόμα υποκλάδος είναι αυτός της διαχείρισης των μέσων μαζικής μεταφοράς. Στη συγκεκριμένη περίπτωση χρειάζεται μελέτη όχι μόνο του οδικού δικτύου αλλά και της συμπεριφοράς των επιβατών. Ο εντοπισμός μοτίβων στις συνήθειες μετακίνησής τους προσφέρει στις αρχές τη δυνατότητα για βελτιστοποίηση τους προγράμματος λειτουργίας αλλά και την καλύτερη αξιοποίηση του στόλου.

Στο [Gura17] παρουσιάζεται ένα σύστημα για την καταγραφή της κίνησης των οχημάτων. Για τον σκοπό αυτό, αντί για GPS γίνεται χρήση των τεχνολογιών *Arduino* [McRo10], *Raspberry Pi*[3], και *ZigBee* [Gisl08], ενώ τα δεδομένα αποθηκεύονται σε μορφή XML σε βάση δεδομένων. Αυτή η επιλογή έχει ως αποτέλεσμα τη μείωση του κόστους υλοποίησης, ταυτόχρονα όμως περιορίζει το χωρικό εύρος λειτουργίας του συστήματος.

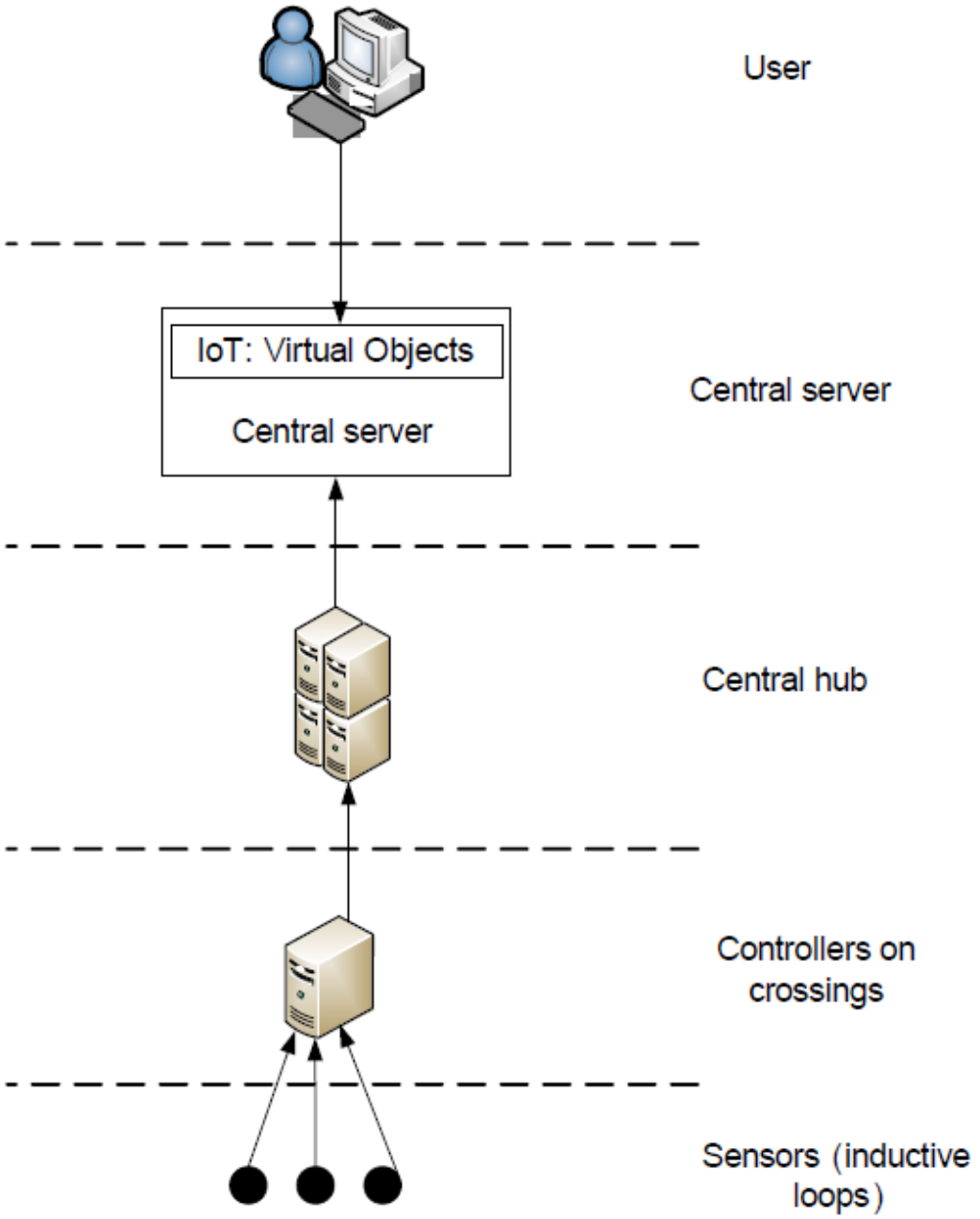
Στο [Ma13] προτείνεται ένα σύστημα για την ανάλυση της συμπεριφοράς των επιβατών στα μέσα μαζικής μεταφοράς, βασισμένο σε στοιχεία που προέρχονται από την πόλη του Πεκίνου. Παρόμοιες προσπάθειες για ανάλογες μελέτες υπήρχαν και κατά το παρελθόν αλλά βασιζόταν κατά κύριο λόγο σε ερωτηματολόγια που συμπλήρωναν οι χρήστες και συνεπώς δεν μπορούσαν να αποτυπώσουν σωστά την πραγματικότητα. Στη συγκεκριμένη περίπτωση αυτό το εμπόδιο δεν υφίσταται καθώς πλέον τα στοιχεία προέρχονται από τη χρήση ηλεκτρονικής κάρτας κατά την επιβίβαση και συνεπώς προσφέρουν σαφήνεια και ακρίβεια ως προς την πληροφορία που μεταφέρουν.

Στο [Ruta18] παρουσιάζεται ένα σύστημα το οποίο λειτουργεί ως βοηθός του οδηγού παρέχοντας πληροφορίες σχετικά με το δρόμο και την κίνηση σε πραγματικό χρόνο κατά τη διάρκεια της μετακίνησης. Οι πληροφορίες αυτές συλλέγονται και ανταλλάσσονται τόσο μεταξύ των οχημάτων σε μια «peer to peer» προσέγγιση όσο και από τον κεντρικό διακομιστή.

[3] `https://www.raspberrypi.org/`

Με αυτό τον τρόπο οι συνθήκες του περιβάλλοντος μπορούν να εκτιμηθούν ανάλογα ώστε να προβλεφθούν πιθανοί κίνδυνοι.

Το [Somo13] ασχολείται με την ανάλυση της κίνησης στους δρόμους της πόλης Enschede της Ολλανδίας. Η υλοποίηση περιλαμβάνει αισθητήρες οι οποίοι λειτουργούν σε μια αρχιτεκτονική πέντε επιπέδων (Σχήμα 2.5). Η καινοτομία στη συγκεκριμένη περίπτωση έγκειται στο γεγονός ότι χρησιμοποιείται μια τεχνική ψηφιακής αναπαράστασης για τους αισθητήρες, οι οποίοι πλέον ονομάζονται *Virtual Objects* (VO). Με αυτό τον τρόπο καθίσταται δυνατός ο εμπλουτισμός τον δεδομένων με πληροφορίες οι οποίες είναι χρήσιμες για την εξόρυξη γνώσης από αυτά.



Σχήμα 2.5: Αρχιτεκτονική πέντε επιπέδων [Somo13]

Στο [Ma15], οι συγγραφείς χρησιμοποιούν τεχνικές *βαθιάς μάθησης* (deep learning) [LeCu15] για την πρόβλεψη της κυκλοφοριακής συμφόρησης. Τα δεδομένα στην συγκεκριμένη περίπτωση συλλέγονται μέσα από GPS που είναι εγκατεστημένα σε ταξί. Μάλιστα για μεγαλύτερη ταχύτητα εκτέλεσης, η ανάλυση γίνεται με τη χρήση καρτών γραφικών μέσω του περιβάλλοντος CUDA της NVIDIA [Nick08]. Τέλος, στο [Yin13] γίνεται παρουσίαση ενός συστήματος διαχείρισης στάθμευσης με τη χρήση αισθητήρων.

2.5 Διαχείριση τροφίμων

Ένα πολύ σημαντικό και ευαίσθητο πεδίο είναι αυτό τις διαχείρισης των τροφίμων. Κατά το παρελθόν έχουν παρουσιαστεί διατροφικά σκάνδαλα τα οποία οφείλονται στην αδυναμία κατάλληλης παρακολούθησης των τροφίμων. Σε προγενέστερες εποχές η διαδικασία αυτή παρουσίαζε κενά καθώς μεγάλο μέρος της βασιζόταν στον ανθρώπινο παράγοντα γεγονός που εισήγαγε μη αντικειμενικά κριτήρια στην εξέλιξη της. Πλέον με το IoT αυτό αλλάζει, καθώς αυτό μπορεί να γίνει με συστηματικό και αυτοματοποιημένο τρόπο. Στο [Han14] παρουσιάζεται ένα μοντέλο εμπιστοσύνης για τρόφιμα. Η διαφορά με παρελθοντικές μελέτες είναι ότι στη συγκεκριμένη οι παράμετροι σχετίζονται αποκλειστικά με ιστορικά δεδομένα που έχουν καταγραφεί από το σύστημα ή από άλλες πηγές και όχι από εκτιμήσεις ειδικών του κλάδου. Επίσης το γεγονός ότι χρησιμοποιεί και παράγοντες που μεταβάλλονται σε πραγματικό χρόνο όπως η τοποθεσία και η ώρα, έχει τη δυνατότητα να προσφέρει στο χρήστη αποτελέσματα που αντιπροσωπεύουν το προϊόν για τη δεδομένη στιγμή και όχι για κάποια προγενέστερη κατάστασή του.

2.6 Θέματα ασφαλείας

Εκτός όμως από τις υπηρεσίες και τις διευκολύνσεις που παρέχει το IoT μαζί τους εισάγονται και κάποιοι περιορισμοί. Αν και η ασφάλεια δεν αποτελεί αυτόνομο πεδίο εφαρμογής, εντούτοις η σημαντικότητα της είναι τέτοια που αξίζει ξεχωριστής αναφοράς. Οι πληροφορίες που κυκλοφορούν στο δίκτυο σε πολλές περιπτώσεις μπορεί να είναι ιδιαίτερα «ευαίσθητες» και να αφορούν θέματα όπως προσωπικά δεδομένα ασθενών, λειτουργίες οργανισμών και άλλα. Όλα αυτά τα δεδομένα θα πρέπει να μπορεί να διασφαλιστεί ότι δεν θα είναι προσβάσιμα σε τρίτους οι οποίοι ενδέχεται να τα εκμεταλλευτούν με κακόβουλο τρόπο. Τα προβλήματα που προκύπτουν είναι άμεσα συνυφασμένα με τη δομή και τη φύση του IoT καθώς οι συσκευές που χρησιμοποιούνται σε αυτό είναι στην πλειονότητα τους περιορισμένων υπολογιστικών δυνατοτήτων. Το γεγονός αυτό έχει σαν αποτέλεσμα τα κλασσικά πρωτόκολλα ασφαλείας και οι τεχνικές κρυπτογραφίας να μην μπορούν να υλοποιηθούν στην συγκεκριμένη περίπτωση και ως εκ τούτου να πρέπει να διερευνηθούν εναλλακτικές λύσεις.

Στο [Outc17] για αυτό το σκοπό προτείνεται η χρήση blockchain [Swan15], το οποίο ταιριάζει με την κατανεμημένη φύση του IoT. Με αυτό τον τρόπο, αναξιόπιστες μεταξύ τους συσκευές μπορούν να αποφασίζουν συλλογικά και με ασφαλή τρόπο σχετικά με τον έλεγχο εισόδου στο σύστημα. Με αυτή την προσέγγιση υπάρχει επίσης το πλεονέκτημα ότι δεν υπάρχει κίνδυνος από επίθεση σε μεμονωμένο σημείο, όπως συμβαίνει στην περίπτωση που υπάρχει κεντρική διαχείριση.

Το [Diro18] πραγματεύεται τον εντοπισμό και την αναγνώριση επιθέσεων σε συστήματα IoT. Αν και συνεχώς εμφανίζονται νέοι τύποι επιθέσεων, αυτοί στην πλειοψηφία τους αποτελούν παραλλαγή κάποιου προϋπάρχοντος. Αυτή η μικρή έστω διαφοροποίηση είναι ικανή να δυσκολέψει σημαντικά την κλασσική μηχανική μάθηση να δώσει αξιόπιστα αποτελέσματα. Για το λόγο αυτό στη συγκεκριμένη έρευνα χρησιμοποιούνται τεχνικές βαθιάς μάθησης που εκτελούνται σε περιβάλλον παράλληλης επεξεργασίας.

Όπως έχουμε δει ως τώρα, τα κινητά τηλέφωνα παίζουν σημαντικό ρόλο σε αρκετά από τα συστήματα που έχουμε παρουσιάσει. Η πλατφόρμα Android[4] που είναι η πιο δημοφιλής αποτελεί στόχο προς εκμετάλλευση από τους δημιουργούς κακόβολου λογισμικού. Όπως γίνεται εύκολα αντιληπτό, η ύπαρξη τέτοιου λογισμικού μπορεί να αποτελέσει απειλή για όλο το σύστημα μέσα στο οποίο λειτουργεί και συνδιαλλάσσεται το εν λόγω τηλέφωνο. Για το σκοπό αυτό, στο [Alam13] χρησιμοποιείται ο αλγόριθμος *τυχαίου δάσους* (random forest) [Ho95] για να μπορέσει να αποφασίσει εάν ένα κινητό είναι «μολυσμένο» ή όχι.

2.7 Αναγνώριση Προσώπου

Η *αναγνώριση προσώπου* (face recognition) [Li11] είναι ένας από τους παλαιότερους και πιο γνωστούς κλάδους με τον οποίο καταπιάνεται η μηχανική μάθηση. Έως τώρα έχουν αναπτυχθεί πολλές τεχνικές για την περάτωσή της, οι οποίες όμως απαιτούν σημαντικούς υπολογιστικούς πόρους. Η εμφάνιση του IoT και η πληθώρα καμερών που αυτό περιλαμβάνει δημιουργεί ένα νέο και πολλά υποσχόμενο πεδίο για τον κλάδο. Το πρόβλημα που προκύπτει στη συγκεκριμένη περίπτωση, όπως έχουμε δει και παραπάνω, είναι ότι τα εν λόγω συστήματα είναι κατά κύριο λόγο χαμηλής υπολογιστικής δυνατότητας. Στο [Amos16a] προτείνεται η βιβλιοθήκη *OpenFace* [Amos16b] που αποτελεί μία προσπάθεια για να καλυφθεί αυτό το κενό. Σκοπός της είναι να μπορεί να δίνει ακριβείς απαντήσεις σε σύντομο διάστημα σε πραγματικό χρόνο, λαμβάνοντας υπόψιν το περιβάλλον στο οποίο βρίσκεται ο χρήστης. Επίσης δίνεται έμφαση ότι, σε αντίθεση με την στατική υλοποίηση, τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση (περιβάλλον) αλλάζουν συνεχώς.

[4] `https://www.android.com/`

Κεφάλαιο 3

Μηχανική Μάθηση

Στο προηγούμενο κεφάλαιο έχουμε ήδη παρουσιάσει το IoT και κάποιες από τις πιο σημαντικές εφαρμογές του. Στις περισσότερες από αυτές, αν όχι σε όλες, βασικό στοιχείο αποτελούν τα δεδομένα διαφόρων τύπων τα οποία συλλέγονται στην εκάστοτε περίπτωση από δομές του δικτύου όπως οι αισθητήρες. Τα δεδομένα αυτά, αν και δύναται να εμπεριέχουν σημαντικές πληροφορίες, στην πρωτογενή τους μορφή αυτή δεν είναι ορατή από την ανθρώπινη διαίσθηση. Αυτό ακριβώς είναι και το σημείο στο οποίο η μηχανική μάθηση σαν εργαλείο καλείται να μετατρέψει απλούς αριθμούς και μετρήσεις σε χρήσιμη γνώση. Για αυτό το λόγο δεν θα ήταν υπερβολή να πούμε ότι χωρίς αυτήν, το IoT δεν θα είχε καν λόγο ύπαρξης.

Με έναν πιο επίσημο ορισμό, μπορούμε να πούμε ότι η μηχανική μάθηση αποτελεί τον επιστημονικό κλάδο που ασχολείται με την έρευνα των αλγορίθμων, οι οποίοι με βάση τα δεδομένα που τους παρέχονται εκπαιδεύουν μοντέλα, τα οποία στη συνέχεια χρησιμοποιούνται για τη διενέργεια προβλέψεων.

Για την υλοποίηση και την επίτευξη σωστών προβλέψεων απαιτούνται μια σειρά ενεργειών από τον χρήστη, τις οποίες θα αναφέρουμε εδώ επιγραμματικά, αλλά θα τις συναντήσουμε και στην πρακτική τους εφαρμογή στη συνέχεια του Κεφαλαίου.

- Τα δεδομένα σε ορισμένες περιπτώσεις προέρχονται από τις πηγές τους σε μορφή ακατάλληλη για επεξεργασία. Για αυτό το λόγο ενδέχεται να χρειαστεί κάποια προεπεξεργασία από το διαχειριστή του συστήματος πριν γίνει η τροφοδότησή τους στον αλγόριθμο.

- Η επιλογή του αλγορίθμου για την εκπαίδευση είναι μια επίσης σημαντική διαδικασία για την οποία πρέπει να συνυπολογιστούν όλοι οι παράμετροι που εμπλέκονται στην λειτουργία του συστήματος. Ταχύτητα εκτέλεσης, διαθέσιμοι υπολογιστικοί πόροι αλλά και ακρίβεια των προβλέψεων είναι ορισμένα κριτήρια που πρέπει να ληφθούν υπόψη για την κατάλληλη επιλογή.

- Επιλογή των μετρικών απόδοσης αλλά και πιθανή εξέταση άλλων αλγορίθμων για σύγκριση των αποτελεσμάτων.

Στη συνέχεια του κεφαλαίου παρουσιάζουμε πρακτικές εφαρμογές των αλγορίθμων ML στο IoT, ταξινομημένους σύμφωνα με τη διεργασία την οποία επιλύουν.

3.1 Ταξινόμηση

Η ταξινόμηση εμπίπτει στην κατηγορία της επιβλεπόμενης μάθησης. Τα δεδομένα εισόδου φέρουν μια ετικέτα η οποία εντάσσει το καθένα από αυτά σε μια συγκεκριμένη κατηγορία. Σκοπός της ταξινόμησης είναι να παράξει ένα μοντέλο το οποίο, για μελλοντικά δεδομένα τα οποία δεν θα φέρουν την αντίστοιχη ετικέτα, να μπορεί να προβλέψει την κατηγορία στην οποία ανήκουν.

Στο [Bove14] προτείνεται μια αρχιτεκτονική για την υλοποίηση αλγορίθμων ML σε κατανεμημένο περιβάλλον με χαμηλής υπολογιστικής ισχύος μονάδες, όπως ακριβώς συμβαίνει στο IoT. Η επικοινωνία των συσκευών γίνεται μέσω RESTful API (HTTP ή CoAP). Για λόγους διευκόλυνσης της επαναχρησιμοποίησης σε διαφορετικά δίκτυα αισθητήρων αλλά και για να αποκρυφθεί η πολυπλοκότητα των δεδομένων, εισάγονται δύο νέες οντότητες, *Virtual Sensor* και *Virtual Class*. Η απόφαση για το πρόβλημα ταξινόμησης λαμβάνεται στον Virtual Sensor όπου συγκεντρώνονται όλες οι πιθανότητες που έχουν υπολογιστεί στις Virtual Classes. Για την διαδικασία αυτή υπάρχουν δύο κατηγορίες λειτουργίας του συστήματος. Στην «End-to-end» δεδομένα από τις Virtual Classes αποστέλλονται μόνο όταν αυτά ζητηθούν από τον Virtual Sensor ενώ στην «Sync-based» αυτό γίνεται κάθε φορά που υπάρχει διαφοροποίηση στην υπολογιζόμενη πιθανότητα, με την δεύτερη να είναι κατάλληλη σε περιπτώσεις που απαιτείται ακρίβεια και γρήγορη απόκριση. Για την εφαρμογή του συστήματος παρουσιάζεται ένα παράδειγμα όπου μέσω της παρακολούθησης της κατανάλωσης ρεύματος, γίνεται πρόβλεψη για τον τύπο αλλά και την κατάσταση της συσκευής η οποία βρίσκεται σε λειτουργία. Ο αλγόριθμος που χρησιμοποιείται είναι τα *κρυφά μαρκοβιανά μοντέλα* (Hidden Markov Models ή HMMs) [Baum66], καθώς από την φύση του ευνοεί τον καταμερισμό των απαραίτητων υπολογισμών στα διαφορετικά μέρη του συστήματος, ενώ η εκπαίδευση του μοντέλου έχει γίνει από πριν με βάση τα διαθέσιμα δεδομένα.

Στο [Han14] παρουσιάζεται ένα μοντέλο εμπιστοσύνης που χρησιμοποιείται για τον έλεγχο τροφίμων.Το γεγονός ότι τα δεδομένα προέρχονται από το IoT και όχι από γενικές παρατηρήσεις και εκτιμήσεις ειδικών, αυξάνει την ακρίβεια και την εγκυρότητα τους. Το πρώτο βήμα που πρέπει να γίνει για την κατασκευή του μοντέλου είναι ο προσδιορισμός των παραγόντων που επηρεάζουν την αξιοπιστία του τροφίμου και στη συνέχεια να ορισθούν οι συναρτήσεις που θα υπολογίζουν την τιμή των παραγόντων αυτών. Για τον υπολογισμό της τελικής τιμής της εκτίμησης εμπιστοσύνης γίνεται χρήση **μπεϋζιανού δικτύου** (bayesian network) [Pear85]. Οι παράγοντες που αναφέραμε πιο πάνω ομαδοποιούνται σε δύο κόμβους ανάλογα με το αν οι τιμές τους είναι μεταβαλλόμενες ή στατικές. Ο διαχωρισμός αυτός είναι χρήσιμος

καθώς διευκολύνει την προσαρμογή των πιθανοτήτων του δικτύου σε περίπτωση που παρατηρείται ότι μια ομάδα παραγόντων επηρεάζει με παρόμοιο τρόπο το αποτέλεσμα, δηλαδή την αξιοπιστία του προϊόντος.

Στο [Pand17] έχοντας ως δεδομένα τους καρδιακούς παλμούς ενός ατόμου, γίνεται ταξινόμηση σχετικά με το αν το συγκεκριμένο άτομο εμφανίζει επίπεδα στρες μεγαλύτερα από το κανονικό. Για το σκοπό αυτό εφαρμόστηκαν αρχικά ο αλγόριθμος VF-15 [Guve97] και ο *απλός μπεϋζιανός ταξινομητής* (naive bayesian classifier) [Maro61], οι οποίοι όμως δεν είχαν ικανοποιητικά αποτελέσματα ως προς την ακρίβεια τους. Στη συνέχεια στα ίδια δεδομένα δοκιμάστηκαν οι αλγόριθμοι *λογιστικής παλινδρόμησης* (Logistic Regression ή LR) [Cram02] και *μηχανών διανυσμάτων υποστήριξης* (Support Vector Machines ή SVM) [Cort95], με τους οποίους επιτευχθεί μεγαλύτερη ακρίβεια.

Στο [Kuma18] χρησιμοποιείται LR, ώστε το σύστημα να μπορεί να αποφανθεί αν ο ασθενής πάσχει από κάποια καρδιακή πάθηση. Η υλοποίηση γίνεται με την βιβλιοθήκη Apache Mahout [Lyub16] σε κατανεμημένο περιβάλλον. Επίσης με ανάλυση καμπύλης *λειτουργικού χαρακτηριστικού δέκτη* (Receiver Operating Characteristic ή ROC) [Fawc06] πάνω στα αποτελέσματα, γίνεται επιπλέον αξιολόγηση σχετικά με το ποιοι παράγοντες είναι πιο σημαντικοί και σχετίζονται με τις καρδιακές παθήσεις οι οποίες πρέπει να προβλεφθούν.

Οι συγγραφείς στο [Lee16] περιγράφουν τη διενέργεια μιας διαδικασίας μηχανικής μάθησης βασισμένης σε *γκαουσιανά μοντέλα μίξης* (Gaussian Mixture Models ή GMMs) [Reyn15], με τη χρήση αλγορίθμου *μεγιστοποίησης αναμονής* (Expectation-Maximization ή EM) [Demp77]. Η ιδιαιτερότητα της εν λόγω έρευνας είναι ότι η υλοποίηση του συστήματος γίνεται πάνω σε ενσωματωμένη πλακέτα. Η πρακτική υλοποίηση που τη συνοδεύει αφορά τον έλεγχο για την πρόβλεψη της κατάστασης του νερού σε ένα καυστήρα. Μάλιστα η ακρίβεια τον αποτελεσμάτων της, αν και πρόκειται για πραγματικού χρόνου εκπαίδευση και ανάλυση ήταν παρόμοια, με την αντίστοιχη προσομοίωσης που έγινε σε περιβάλλον MATLAB[1].

Όπως έχει ήδη αναφερθεί, τα δεδομένα που σχετίζονται με τις εφαρμογές του IoT χαρακτηρίζονται πολλές φορές από το μεγάλο όγκο τους αλλά και την ετερογένεια τους. Για αυτό το σκοπό, στο [Li18] προτείνεται ένα *βαθύ συνελικτικό δίκτυο* (deep convolutional network) [LeCu15] για ταξινόμηση μεγάλων δεδομένων.

Στο [Alam13] μελετάται το κατά πόσο είναι εφικτό να γίνει ταξινόμηση για κινητά τηλέφωνα με λογισμικό Android ως προς το αν έχουν μολυνθεί με κακόβουλο λογισμικό με τη χρήση του αλγόριθμου τυχαίου δάσους [Ho95]. Αφού διαπιστώθηκε ότι επιτυγχάνεται υψηλή ακρίβεια σε σύγκριση και με άλλους αλγορίθμους διερευνήθηκε και το πόσο επηρεάζεται η επίδοση από τις παραμέτρους που χρησιμοποιούνται. Με βάση τα πειράματα παρατηρήθηκε ότι τα η αύξηση του αριθμού των δέντρων βελτιώνει τα αποτελέσματα ενώ προσδιορίζεται και το απαραίτητο βάθος που πρέπει να έχουν αυτά. Το μειονέκτημα αυτής της διαδικασίας

[1] `https://www.mathworks.com/products/matlab.html`

είναι ότι δεν εκτελείται σε πραγματικό χρόνο και έτσι δεν μπορεί να εξασφαλίσει συνεχή προστασία για την συσκευή.

Στο [Hoss16] χρησιμοποιείται ο αλγόριθμος *one-class SVM* (OCSVM) [Scho99] για να ελέγξει τους καρδιακούς παλμούς ενός ατόμου και να αποφανθεί αν αυτοί είναι φυσιολογικοί ή ενδέχεται να υπάρχει κάποια καρδιακή δυσλειτουργία. Η διαφορά με τον κλασσικό αλγόριθμο SVM είναι ότι στον OCSVM υπάρχει μόνο μία κλάση στην εκπαίδευση αντί για δύο.

Το [Khan14] καταπιάνεται με την ταξινόμηση σε *ροές δεδομένων* (data streams). Όπως έχουμε ήδη επισημάνει, οι αισθητήρες αλλά και όλες οι συσκευές που αποτελούν το IoT παράγουν δεδομένα με τόσο υψηλό ρυθμό που καθιστά δύσκολη την διαχείρισή τους τόσο από την πλευρά της ανάλυσής τους όσο και της αποθήκευσης τους. Γίνεται λοιπόν εύκολα κατανοητό ότι για την αποτελεσματική επεξεργασία τους χρειάζεται ειδική προσέγγιση. Το πρώτο βήμα που πραγματοποιείται στη συγκεκριμένη εργασία είναι η εφαρμογή ενός αλγορίθμου *απλής προσθετικής προσέγγισης* (simple aggregate approximation ή SAX) στις χρονοσειρές δεδομένων με σκοπό να μειωθούν οι διαστάσεις τους. Θα πρέπει βέβαια να επισημανθεί ότι, ανάλογα με την περίπτωση, πριν από αυτό θα πρέπει να έχουν αντιμετωπισθεί προβλήματα που αφορούν δεδομένα που λείπουν με τη διαγραφή των αντίστοιχων εγγραφών. Η επόμενη κίνηση επιβάλλει την εύρεση των πιθανών κλάσεων-στόχων για το σύστημα. Αυτό επιτυγχάνεται με την εφαρμογή αλγορίθμου συσταδοποίησης πάνω στα δεδομένα. Ο αλγόριθμος που επιλέγεται στην προκειμένη περίπτωση είναι ο DBScan [Este96], καθώς είναι ιδιαίτερα βολικό το γεγονός ότι δεν απαιτεί την εκ των προτέρων γνώση του αριθμού των κλάσεων που αναζητούμε, κάτι που συνήθως ισχύει όταν μελετάμε ροές δεδομένων. Τα αποτελέσματα που προκύπτουν από αυτή τη διαδικασία ταξινομούνται με τη βοήθεια του αλγορίθμου SVM για να προκύψει το τελικό ζητούμενο. Η ακρίβεια που επιτυγχάνεται φτάνει το 80%, αλλά το γεγονός ότι για την ανάλυση χρειάζεται να διαβαστούν δύο φορές τα δεδομένα αυξάνει σημαντικά τον υπολογιστικό χρόνο που απαιτείται για την ολοκλήρωση του.

Στο [Uzel15] παρουσιάζεται μια μέθοδος για την εκτίμηση του επιπέδου προσοχής των μαθητών κατά τη διάρκεια μιας διάλεξης. Τα δεδομένα έχουν συλλεχθεί με τη βοήθεια αισθητήρων και παρουσιάζουν την κατάσταση του περιβάλλοντος (θερμοκρασία, υγρασία, επίπεδα θορύβου) και τα ανάλογα χρονικά διαστήματα έχουν χαρακτηρισθεί από τους ίδιους τους μαθητές ως διάστημα με συγκέντρωση ή χωρίς. Στη συνέχεια τα δεδομένα χρησιμοποιήθηκαν για την εκπαίδευση δέκα διαφορετικών αλγορίθμων ταξινόμησης. Τελικά αυτός με το καλύτερο ποσοστό αναγνώρισης αποδείχτηκε ο AdaBoost M1 [Freu96]. Το γεγονός αυτό δικαιολογείται από το ότι ο συγκεκριμένος αλγόριθμος είναι *υψηλής μεροληψίας* (high bias) και *χαμηλής διακύμανσης* (low variance) και συνεπώς ευνοείται από το μικρό μέγεθος του συνόλου δεδομένων της μελέτης.

3.2 Συσταδοποίηση

Η συσταδοποίηση εμπίπτει στην κατηγορία της μη επιβλεπόμενης μάθησης. Σε αντίθεση με την ταξινόμηση, τα δεδομένα που εισάγονται δεν φέρουν μαζί τους κάποια ετικέτα που να τα εντάσσει σε κάποια κατηγορία. Σκοπός της συσταδοποίησης είναι να εντοπίσει της κατηγορίες αυτές έτσι ώστε να μπορεί να κατατάξει τα μελλοντικά δεδομένα. Ανάλογα με την περίπτωση, το πλήθος των κατηγοριών μπορεί να είναι γνωστό από πριν ή όχι.

Στο [Hrom15] οι ερευνητές επεξεργάζονται δεδομένα σχετικά με το περιβάλλον και την ατμοσφαιρική ρύπανση στην πόλη του Ζάγκρεμπ. Η συλλογή των απαραίτητων πληροφοριών γίνεται με τη συνδρομή πολιτών οι οποίοι κατά τη διάρκεια των μετακινήσεών τους στην πόλη φέρουν ειδικούς αισθητήρες, ικανούς να καταγράφουν δείκτες όπως η θερμοκρασία, η υγρασία αλλά και τα επίπεδα ρύπων. Με δεδομένο όμως ότι τα μεγάλα αστικά κέντρα παρουσιάζουν ανομοιογένεια ως προς τον τρόπο ανάπτυξης τους υπάρχει η ανάγκη να γίνει διαχωρισμός των στοιχείων ανάλογα με την περιοχή από την οποία προέρχονται, ώστε η περαιτέρω ανάλυση που θα γίνει σε αυτά να είναι περισσότερο στοχευμένη. Στη συγκεκριμένη περίπτωση και λαμβάνοντας υπόψη τη δομή της πόλης οι συγγραφείς την έχουν χωρίσει σε 4 τμήματα καθένα από τα οποία χαρακτηρίζεται από το κέντρο του. Αφού λοιπόν τα κέντρα είναι γνωστά εκ των προτέρων, ο αλγόριθμος k-μέσων [MacQ67] αποτελεί ιδανική επιλογή για να πραγματοποιήσει τον διαχωρισμό των σημείων σε συστάδες.

Στο [Ma13] παρουσιάζεται μια έρευνα σχετικά με τις συνήθειες των πολιτών κατά τη χρήση των μέσω μαζικής μεταφοράς. Τα δεδομένα για την ανάλυση προέρχονται το σύστημα «έξυπνων καρτών» που χρησιμοποιούν τα λεωφορεία στο Πεκίνο για τις πληρωμές των διαδρομών. Το γεγονός ότι το σύστημα αυτό δεν είχε σχεδιασθεί με πρόβλεψη για πιθανές μελλοντικές αναλύσεις όπως αυτή, τα δεδομένα από μόνα τους είναι σε ορισμένες περιπτώσεις ελλειπή και πρέπει να συνδυαστούν μεταξύ τους ώστε να προκύψει η ζητούμενη πληροφορία. Μέσα από αυτή τη διαδικασία προκύπτουν και οι αλυσίδες ταξιδιών των επιβατών οι οποίες με τη χρήση του αλγόριθμου DBScan [Este96] μας δίνουν το ιστορικό μοτίβο μετακινήσεων. Σημαντικό κριτήριο για την επιλογή του συγκεκριμένου αλγορίθμου ήταν το γεγονός ότι για την εφαρμογή του δεν απαιτεί τη γνώση του αριθμού των συστάδων που υπάρχουν. Σε δεύτερο στάδιο εξετάζεται πόσο τακτικό είναι ο κάθε επιβάτης ως προς τη χρήση των μέσων. Για το σκοπό αυτό χρησιμοποιούνται οι αλγόριθμοι k-μέσων++ [Arth07] και rough-set theory [Pawl98], οι οποίοι επιλέχθηκαν λόγω καλύτερου συνδυασμού ακρίβειας και χρονικής πολυπλοκότητας.

Στο [Ganz15] εξετάζονται 2 παραδείγματα όπου και στα 2 γίνεται χρήση του αλγορίθμου k-μέσων [MacQ67]. Η πρώτη περίπτωση αφορά την παρατήρηση της κατανάλωσης ηλεκτρικού ρεύματος σε περιβάλλον γραφείου. Ο αλγόριθμος εκτελείται με παράμετρο $k = 2$ όσες είναι και οι συστάδες που θέλουμε να έχουμε, μία για τις εργάσιμες ημέρες και μια για τις

μη-εργάσιμες, οπότε και η κατανάλωση είναι χαμηλότερη. Στην δεύτερη περίπτωση γίνεται παρακολούθηση των καρδιακών παλμών ενός ασθενή. Εδώ αντίστοιχα η εκτέλεση γίνεται με παράμετρο $k = 3$, με τις συστάδες να αντιστοιχούν σε καταστάσεις για χαμηλή και υψηλή καρδιακή λειτουργία, αλλά και για μετρήσεις που είναι εκτός φυσιολογικών ορίων, όπου είτε πρόκειται για κάποιο σφάλμα κατά τη μέτρηση είτε υπάρχει πιθανό πρόβλημα για τον ασθενή.

Όπως έχει ήδη αναφερθεί, οι ροές δεδομένων, λόγω του μεγάλου όγκου δεδομένων που περιλαμβάνουν ,αλλά και την συνεχή ροή αυτών απαιτούν στις περισσότερες των περιπτώσεων διαφορετική προσέγγιση, σε σχέση με τα παραδοσιακά σύνολα δεδομένων. Το [Yogi13] αποτελεί μια έρευνα που παρουσιάζει κάποιους από τους αλγορίθμους που χρησιμοποιούνται για συσταδοποίηση πάνω σε ροές δεδομένων. Οι αλγόριθμοι που παρουσιάζονται χωρίζονται σε δύο κατηγορίες με βάση τον τρόπο με τον οποίο αντιμετωπίζουν τα δεδομένα. Η πρώτη κατηγορία που ονομάζεται *συσταδοποίηση κατά παράδειγμα* (clustering by example) μεταχειρίζεται τα σημειακά δεδομένα που έρχονται την ίδια χρονική στιγμή από διαφορετικές πηγές ως μία μονάδα. Κάθε μονάδα δεδομένου περιγράφει τις ιδιότητες μιας οντότητας σε ενα συγκεκριμένο χρονικό σημείο. Αλγόριθμοι που λειτουργούν κατά αυτό τον τρόπο είναι οι STREAM, CluStream, HPStream, DenStream, D-stream, E-Stream, DUCstream, DD-Stream και HUE-Stream. Αντίστοιχα η δεύτερη κατηγορία που ονομάζεται *συσταδοποίηση κατά μεταβλητή* (clustering by variable) μεταχειρίζεται τα σημειακά δεδομένα που έρχονται από μία μεμονωμένη πηγή σαν ροή από δεδομένα και όλα τα σημειακά δεδομένα από την ίδια ροή πρέπει να ανήκουν στην ίδια συστάδα. Τέτοιοι αλγόριθμοι είναι οι Online Divisive-Agglomerative Clustering, Probability and Distribution-based Clustering, COMET-CORE και SPE-Cluster.

3.3 Ανίχνευση Έκτοπων Τιμών

Η διαδικασία της *ανίχνευσης έκτοπων τιμών* (outlier detection) έχει σκοπό τον εντοπισμό μη φυσιολογικών τιμών στα δεδομένα. Τα σωστά αποτελέσματα της μπορεί να είναι ιδιαίτερα κρίσιμα σε περιπτώσεις που χρησιμοποιείται για εντοπισμό προβλημάτων ασφαλείας. Για την επίτευξη της μπορεί να χρησιμοποιείται συνδυασμός τεχνικών που έχουμε δει παραπάνω.

Στο [Jakk10] παρουσιάζεται η εργασία για τον εντοπισμό μη φυσιολογικών τιμών στην κατανάλωση ρεύματος σε οικιακό περιβάλλον. Η πρώτη μέθοδος περιλαμβάνει τη χρήση στατιστικής προσέγγισης. Η δεύτερη εκτελεί συσταδοποίηση ώστε να χωρίσει τα δεδομένα σε φυσιολογικά και μη φυσιολογικά με εφαρμογή του αλγορίθμου *k-πλησιέστερων γειτόνων* (kNN)[Altm92]. Η στατιστική μέθοδος είχε ικανοποιητικά αποτελέσματα όταν ο αριθμός των μη φυσιολογικών τιμών στο dataset ήταν μικρός ενώ σε μεγαλύτερο αριθμό είχε πρόβλημα. Εν αντιθέσει ο αλγόριθμος kNN απέδωσε σταθερά ικανοποιητικά αποτελέσματα χω-

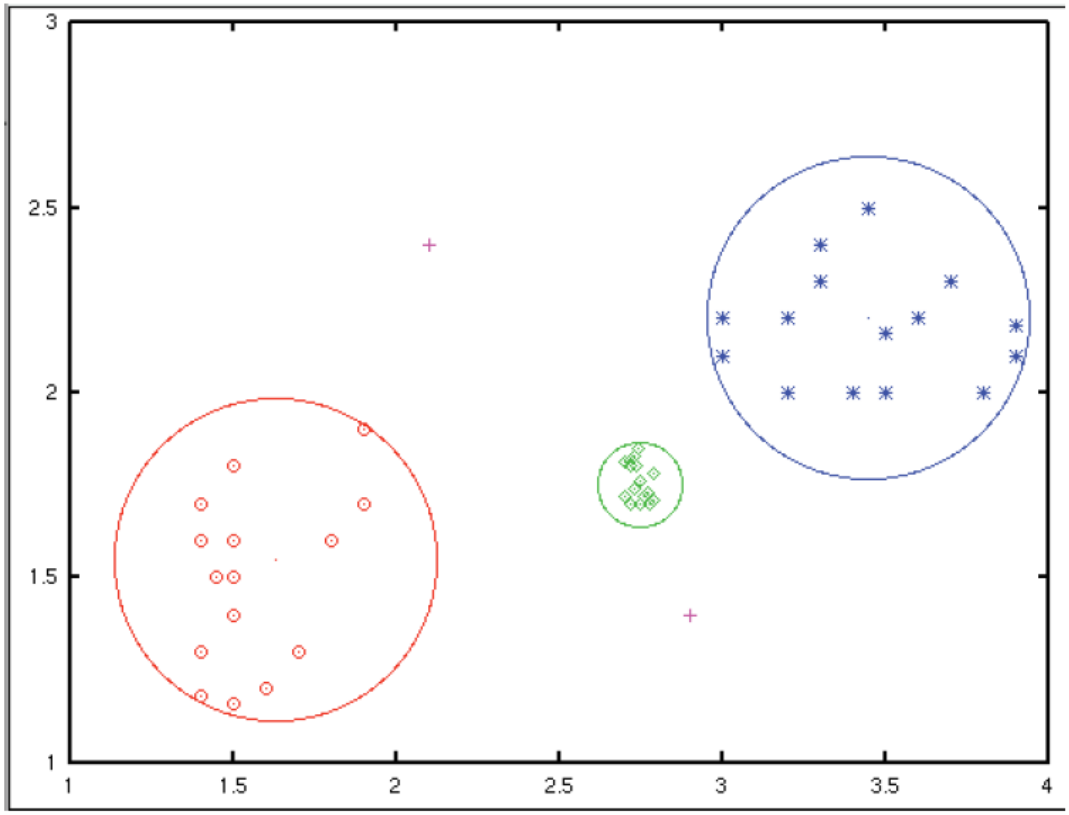
ρίς να επηρεάζεται από αυτόν τον παράγοντα και συνεπώς προτιμάται για τη συγκεκριμένη διαδικασία.

Το [Shuk15] πραγματεύεται την ανίχνευση έκτοπων τιμών πάνω σε ροές δεδομένων. Το γεγονός ότι τα δεδομένα που προέρχονται από ροές σε αντίθεση με τα στατικά έχουν εξάρτηση από τον χρόνο, δημιουργεί την ανάγκη για διαφορετική αντιμετώπιση τους. Ανάλογα με τη χρονική στιγμή στην οποία αναφέρεται ένα δεδομένο, το περιβάλλον είναι διαφορετικό και συνεπώς αλλάζει και η σημασία του ίδιου του δεδομένου. Επίσης όταν υπάρχουν περισσότερες από μία πηγές, τα δεδομένα τους μπορεί να σχετίζονται μεταξύ τους, χωρίς όμως να καταφθάνουν για επεξεργασία την ίδια στιγμή, ενώ οι ροές μπορούν να θεωρηθούν άπειρες αφού πρακτικά μπορούν να να παράγουν συνεχώς δεδομένα. Ένα ακόμα πρόβλημα είναι ότι τα δεδομένα λόγω του τρόπου που δημιουργούνται είναι δύσκολο να έχουν επισημείωση, συνεπώς θα πρέπει να χρησιμοποιούνται τεχνικές μη-επιβλεπόμενης μηχανικής μάθησης. Μια υβριδική μέθοδος για την επίλυση του προβλήματος περιλαμβάνει εφαρμογή του αλγορίθμου DBScan [Este96] για την πρώτη δημιουργία της συστάδας ενώ μετά με χρήση του αλγορίθμου k-μέσων [MacQ67] ανανεώνονται τα βάρη τα οποία υποδεικνύουν ποιες ιδιότητες είναι περισσότερο σχετικές με την ομαδοποίηση και ποιες όχι.

Στο [Souz15] υλοποιείται η διαδικασία ανίχνευσης έκτοπων τιμών σε κατανεμημένο περιβάλλον για μεγάλα δεδομένα με τη χρήση των Apache Hadoop [Vavi13] και Mahout [Lyub16]. Πάνω στα αρχικά δεδομένα εφαρμόζεται ο αλγόριθμος συσταδοποίησης canopy [McCa00], ο οποίος παράγει ως αποτέλεσμα τον προτεινόμενο αριθμό συστάδων που προκύπτει ως πρόβλεψη. Η αρχική εκτίμηση για τον αριθμό και τα κέντρα των συστάδων εισάγεται στη συνέχεια ως είσοδος στον αλγόριθμο k-μέσων [MacQ67], ο οποίος υπολογίζει το τελικό αποτέλεσμα της ομαδοποίησης που προτείνεται. Έχοντας πλέον υπολογίσει το κέντρο και την ακτίνα όλων των συστάδων, για να εντοπισθεί αν ένα σημείο αποτελεί μη προβλεπόμενη τιμή υπολογίζεται η τιμή της απόστασης του από κάθε κέντρο και αν αυτή είναι μεγαλύτερη από κάθε ακτίνα τότε πρόκειται για τέτοια. Παράδειγμα αυτής της διαδικασίας φαίνεται στο Σχήμα 3.1.

3.4 Συσχέτιση Δεδομένων

Το IoT αποτελεί ένα δίκτυο στο οποίο συνδέονται συσκευές και αισθητήρες που παράγουν δεδομένα διαφορετικών τύπων. Εξετάζοντας μεμονωμένα τον κάθε τύπο τα δεδομένα που έχουμε είναι πιθανό να δώσουν μόνο ένα μέρος της χρήσιμης πληροφορίας που εμπεριέχουν. Για αυτό το σκοπό είναι χρήσιμο να μελετηθούν και οι συσχετίσεις που έχουν μεταξύ τους τα διαφορετικά δεδομένα, τόσο χωρικά όσο και χρονικά. Για παράδειγμα, η θερμοκρασία σε μία περιοχή μπορεί να αποτελεί παράγοντα που επηρεάζει την υγρασία στην ίδια περιοχή σε μεταγενέστερη χρονική στιγμή. Αυτή τη γνώση μπορούμε να την αποκτήσουμε

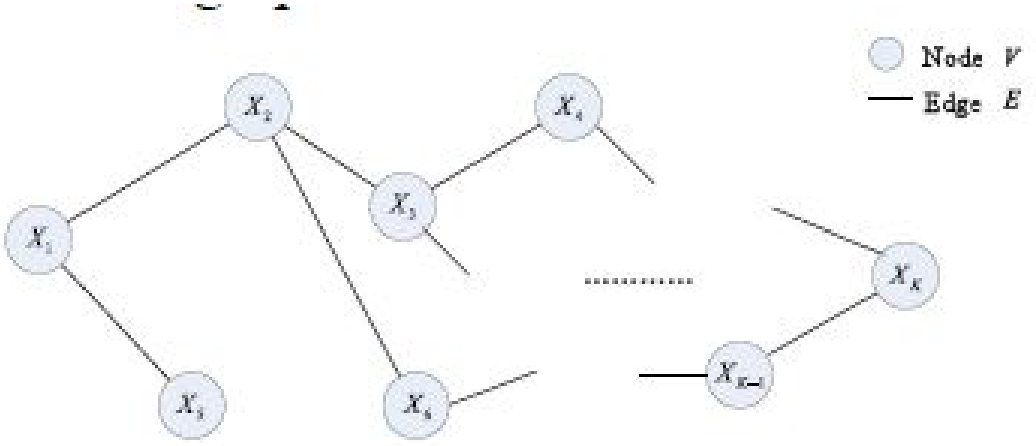


Σχήμα 3.1: Γραφική απεικόνιση έκτοπων τιμών [Souz15]

μελετώντας τα διαθέσιμα δεδομένα στο σύνολό τους. Στο [Dong11] υλοποιούν αυτή τη διαδικασία για να εξετάσουν τη συσχέτιση μεταξύ θερμοκρασίας, υγρασίας και φωτεινότητας με τη χρήση *γκαουσιανών γραφικών μοντέλων* [Hojs12] όπως φαίνεται και στο Σχήμα 3.2).

Στο [Hrom15] διερευνάται και εκεί η συσχέτιση μεταξύ περιβαλλοντικών παραγόντων. Για το σκοπό αυτό παρουσιάζουν γράφους συσχέτισης όπου κάθε κόμβος αντιστοιχεί σε ένα παράγοντα και οι ακμές μεταξύ των κόμβων εκφράζουν τη συσχέτιση τους. Με δεδομένο ότι από την βιβλιογραφία προτείνεται γραμμική συσχέτιση για αυτούς τους παράγοντες, η έρευνα επεκτείνεται με αυτό το δεδομένο. Για αυτό το σκοπό χρησιμοποιείται ο συντελεστής συσχέτισης Pearson, ο οποίος δίνει τιμές στο διάστημα [-1,1] με τα δύο άκρα να δηλώνουν πλήρη κανονική ή ανάστροφη γραμμική συσχέτιση. Βέβαια θα πρέπει να σημειωθεί ότι στην περίπτωση που δεν προκύψει γραμμική συσχέτιση (όταν έχουμε τιμή 0), αυτό δεν σημαίνει ότι δεν υπάρχει κάποιου άλλου είδους συσχέτιση μεταξύ των παραγόντων.

Όταν υπάρχει κάποιο δεδομένο με τιμή εκτός από τα φυσιολογικά όρια τότε υπάρχουν δύο πιθανά ενδεχόμενα. Το πρώτο είναι να αποτελεί απλά εξαίρεση στην κανονικότητα του αντικειμένου που παρακολουθείται και το δεύτερο να υπάρχει σφάλμα στην μέτρηση του αισθητήρα. Το [Mone13] προτείνει μια λύση σε αυτό το πρόβλημα με προσπάθεια συσχέτισης μεταξύ των αισθητήρων που συνυπάρχουν σε ένα δίκτυο. Με αυτό τον τρόπο γίνεται εφικτό



Σχήμα 3.2: Απεικόνισης γκαουσιανών γραφικών μοντέλων [Dong11]

να ελεγχθεί το αν οι τιμές που λαμβάνονται ως είσοδοι είναι οι πραγματικές. Για τον σκοπό αυτό γίνεται χρήση της *ανάλυσης κανονικής συσχέτισης* [HOTE36] και της *ανάλυσης κυρίων συνιστωσών* [FRS01].

3.5 Ανάλυση Χρονοσειρών

Ένας από τους βασικούς πυλώνες πάνω στους οποίου στηρίζεται το IoT είναι οι αισθητήρες. Οι τιμές που αυτοί καταγράφουν σε βάθος χρόνου μπορούν να πάρουν τη μορφή χρονοσειρών. Η ανάλυση αυτών μπορεί να εξάγει χρήσιμα στοιχεία αλλά και προβλέψεις για την εξέλιξη των δεικτών που απεικονίζουν στο μέλλον. Στο [Ni14] παρουσιάζεται μια προσέγγιση για την επίτευξη αυτού του σκοπού. Τα δεδομένα εισόδου σε πρώτο στάδιο υφίστανται προ-επεξεργασία με την μέθοδο Ensemble Empirical Mode Decomposition (EEMD) [WU09]. Με αυτό τον τρόπο τα δεδομένα διασπώνται σε επιμέρους χρονοσειρές και αποστέλλονται στο κομμάτι της πρόβλεψης όπου κάθε σειρά λογίζεται ως δεδομένα εκπαίδευσης για το αντίστοιχο μοντέλο *παλινδρόμησης διανυσμάτων υποστήριξης* (Support Vector Regression ή SVR) [Druc97]. Κατά τη διάρκεια δημιουργία του κάθε μοντέλου SVR, χρησιμοποιείται ο αλγόριθμος Particle Swarm Optimization (PSO) [Kenn10], με σκοπό την βελτιστοποίηση των παραμέτρων του.

Κεφάλαιο 4

Δεδομένα και Μετρικές

4.1 Δεδομένα

Στο παρόν κεφάλαιο θα ασχοληθούμε με τα δεδομένα τα οποία χρησιμοποιούνται στις εφαρμογές του IoT, τις οποίες έχουμε μελετήσει έως τώρα. Έχει γίνει ήδη αντιληπτή η σπουδαιότητα αυτών καθώς και ο άμεσος επηρεασμός τόσο στον σχεδιασμό όσο και στη λειτουργία της εκάστοτε εφαρμογής. Αν και η διαχείριση και η επεξεργασία των δεδομένων αποτελεί από μόνη της ένα αυτοτελή κλάδο έρευνας, το IoT μέσα από την ανάπτυξή του εγείρει και αυτό με τη σειρά του ζητήματα και προβληματισμούς που εμπίπτουν στο συγκεκριμένο πεδίο. Στη συνέχεια παρουσιάζουμε ορισμένα από τα βασικά προβλήματα που προκύπτουν καθώς και τη μορφή των δεδομένων που συναντήσαμε στις βιβλιογραφικές αναφορές.

Στην ιδεατή περίπτωση, το IoT θέλει να να αντιλαμβάνεται τον πραγματικό κόσμο ως μια ενιαία οντότητα. Κάτι τέτοιο είναι εύκολα εφικτό σε εφαρμογές περιορισμένης κλίμακας, όπου οι πηγές που παράγουν τα δεδομένα είναι και αυτές περιορισμένες και εύκολα ελέγξιμες από τον ανθρώπινο παράγοντα ώστε να παράγουν ομοιογενή αποτελέσματα. Σε τέτοιες περιπτώσεις η επεξεργασία μπορεί να γίνει άμεσα και με ήδη γνωστές μεθόδους χωρίς να απαιτείται επιπλέον κόπος για προεπεξεργασία και μετασχηματισμό. Η πρακτική δυσκολία παρουσιάζεται όταν οι πηγές αυξάνουν σε αριθμό και τα αποτελέσματα που παράγουν δεν έρχονται όλα στην ίδια μορφή. Η διαφοροποίηση αυτή μπορεί να έγκειται σε συνοδευτικά στοιχεία που μπορεί να υπάρχουν στην εκάστοτε περίπτωση ή σε διαφορετική κλίμακα μέτρησης, στοιχεία που μπορούν να γίνουν εύκολα αντιληπτά από τον ανθρώπινο παράγοντα όχι όμως και από έναν υπολογιστή. Η ετερογένεια αυτή για να αντιμετωπιστεί απαιτεί τη χρήση μεθόδων πέρα των συνηθισμένων. Στο σημείο αυτό βλέπουμε ότι η μηχανική μάθηση (που είναι το έτερο στοιχείο που μελετάμε στη συγκεκριμένη εργασία) χρησιμεύει όχι μόνο στην επεξεργασία των δεδομένων για την παραγωγή του τελικού αποτελέσματος αλλά και στην προεπεξεργασία αυτών ώστε να μετατραπούν σε μορφή με την οποία να μπορούμε να εργαστούμε μαζί τους.

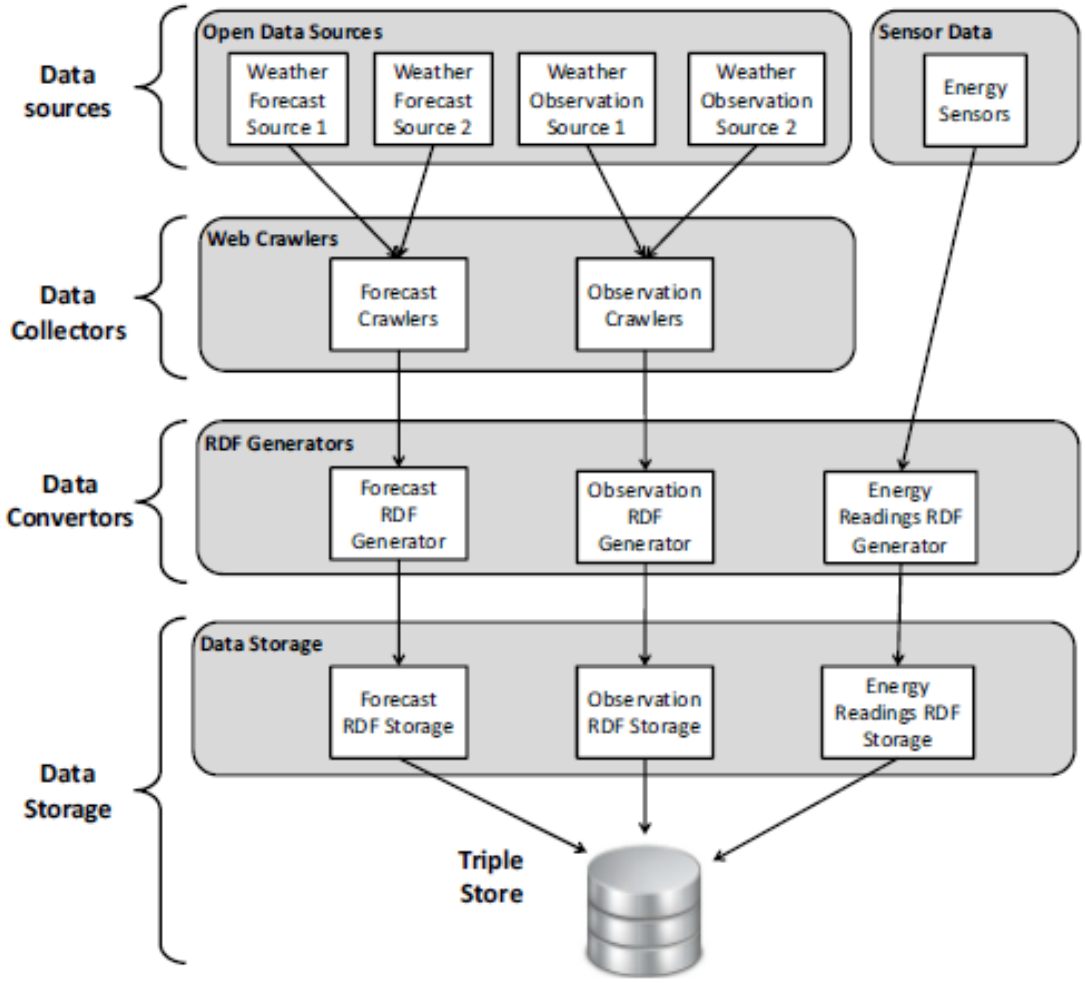
Ένα ακόμα στοιχείο με το οποίο ερχόμαστε αντιμέτωποι είναι ο μεγάλος όγκος των παραγομένων δεδομένων. Η πληθώρα των πηγών αλλά και των στοιχείων που πρέπει να κατα-

γράψουν δημιουργούν ένα σύνολο δυσκόλως διαχειρίσιμο με συμβατικούς τρόπους. Μάλιστα εκτός από τον μεγάλο όγκο που από μόνος του μπορεί να αποτελεί σημαντικό πρόβλημα, ένα ακόμα ζήτημα που προκείπτει σε ορισμένες περιπτώσεις είναι ότι τα δεδομένα μπορεί να βρίσκονται αποθηκευμένα σε διαφορετικά σημεία και κατ' επέκταση αυτό να δυσχεραίνει την απαραίτητη μεταφορά τους κατά το στάδιο της επεξεργασίας. Μια λύση που παρατηρούμε υλοποιείται σε αρκετές μελέτες είναι η χρήση συστημάτων *κατανεμημένου υπολογισμού* (distributed computing) [Pele00]. Ο κατανεμημένος υπολογισμός είναι κι αυτός ένας κλάδος που έχει γνωρίσει μεγάλη εξέλιξη τα τελευταία χρόνια, τόσο σε ερευνητικό όσο και εμπορικό επίπεδο και βρίσκουν πολύ χρήσιμη εφαρμογή στο IoT. Με αυτό τον τρόπο τα δεδομένα και οι εργασίες κατανέμονται, μειώνοντας το συνολικό χρόνο περάτωσης της εργασίας αλλά και την κίνηση στο δίκτυο.

Υπαρκτό είναι επίσης το ενδεχόμενο τα δεδομένα που καλούμαστε να διαχειριστούμε να μην είναι στατικά, αλλά να προέρχονται από ροή δεδομένων. Σε μια τέτοια περίπτωση δεν είναι δυνατή η ύπαρξη συνολικής εικόνας για τα δεδομένα τα οποία μπορούν να θεωρηθούν άπειρα, εγείροντας ταυτόχρονα δυσκολίες σε ότι έχει να κάνει με την αποθήκευση τους. Αποτέλεσμα των παραπάνω είναι ότι ενδέχεται η επεξεργασία να πρέπει να γίνει με την πραγματοποίηση μόνο ενός περάσματος ανάγνωσης. Με την τόσο έντονη εισαγωγή της έννοιας του χρόνου στην επεξεργασία εισάγεται και η ανάγκη για ξεκαθάρισμα του διαστήματος του οποίου τα εκάστοτε δεδομένα παραμένουν έγκυρα και παρέχουν χρήσιμη πληροφορία στο σύστημα. Το ίδιο ισχύει και με την χωρική τους κατανομή, η οποία μπορεί να αλλάζει δυναμικά την σημασία τους και τον τρόπο με τον οποίο θα πρέπει να αξιολογηθούν. Τέλος, μέσα σε αυτό το πλαίσιο δύναται να υπάρχουν πολλαπλές ροές δεδομένων οι οποίες τροφοδοτούν το σύστημα και οι οποίες πρέπει να βρίσκονται σε συγχρονισμό ώστε τα δεδομένα να συνδυάζονται με τον κατάλληλο τρόπο. Αφού τονίσαμε σε θεωρητικό επίπεδο κάποιες από τις προκλήσεις και ιδιαιτερότητες που παρουσιάζουν τα δεδομένα στο IoT, μπορούμε σε αυτό το σημείο να προχωρήσουμε στην παρουσίαση αυτών που συναντήσαμε στις εφαρμογές με τις οποίες ασχοληθήκαμε.

Στο [Derg14] παρουσιάζεται ένα σύστημα για την πρόβλεψη της κατανάλωσης ενέργειας σε κτίριο. Για το σκοπό αυτό αντλούνται από το Διαδίκτυο δεδομένα από διαφορετικές πηγές σχετικά με την πρόβλεψη και την παρατήρηση των καιρικών συνθηκών. Για την επιλογή των καταλλήλων πηγών χρησιμοποιούνται τεχνικές που έχουμε παρουσιάσει στα προηγούμενα κεφάλαια. Παράλληλα μέσα στο κτίριο υπάρχουν αισθητήρες οι οποίοι καταγράφουν και αποθηκεύουν την πορεία κατανάλωσης ηλεκτρικής ενέργειας. Η αρχιτεκτονική για τη διαχείριση των δεδομένων στη συγκεκριμένη εφαρμογή φαίνεται στο Σχήμα 4.1.

Οι συγγραφείς στο [Gura17] παρουσιάζουν μία πρόταση για τον έλεγχο των τεκταινομένων στο δρόμο. Τα οχήματα τα οποία φέρουν τον κατάλληλο εξοπλισμό καταγράφουν στοιχεία που αφορούν τα ίδια αλλά και το περιβάλλον το οποίο συναντούν (π.χ κυκλοφο-



Σχήμα 4.1: Αρχιτεκτονική διαχείρισης δεδομένων [Derg14]

ριακή συμφόρηση) και ανταλλάσσουν αυτές τις πληροφορίες μεταξύ τους. Τα στοιχεία καταγράφονται και κεντρικά σε μία βάση XML από όπου είναι ευκολότερο να επεξεργαστούν. Βλέπουμε λοιπόν ότι στη συγκεκριμένη εργασία υπάρχει συνδυασμός σχετικά με τον τρόπο που μοιράζονται και αξιοποιούνται τα δεδομένα.

Στο [Han14] προτείνεται ένα μοντέλο αξιοπιστίας για τρόφιμα. Για την επίτευξη του χρησιμοποιούνται δεδομένα σχετικά με το κάθε προϊόν τα οποία χωρίζονται σε μεταβαλλόμενα και στατικά. Στην πρώτη κατηγορία περιλαμβάνονται ο χρόνος παραμονής στο ράφι, η τοποθεσία καθώς και η υπογραφή που βοηθάει στην πιστοποίηση της γνησιότητας του προϊόντος. Στη δεύτερη συναντάμε την αλυσίδα παραγωγής όπου βλέπουμε τους εμπλεκόμενους στην πορεία του προϊόντος (όσο περισσότεροι τόσο περισσότερες και οι πιθανότητες για την εμφάνιση σφάλματος), τον ρυθμό κάλυψης και την φήμη της εταιρίας παραγωγής. Τα δεδομένα αυτά αποτελούν την είσοδο για το μοντέλο το οποίο αποφασίζει για το αν ένα προϊόν μπορεί να θεωρηθεί αξιόπιστο.

Το [Hrom15] χρησιμοποιεί δεδομένα που προέρχονται από αισθητήρες ενσωματωμένους στα ρούχα και τα οχήματα πολιτών. Αυτά περιέχουν μετρήσεις για τη θερμοκρασία, την υγρασία, την ατμοσφαιρική πίεση και τους ρύπους συνοδευόμενες από τον χωροχρονικό προσδιορισμό τους. Με αυτό τον τρόπο είναι δυνατή η μελέτη του συσχετισμού των διαφόρων παραγόντων σε συνάρτηση τόσο με το χώρο όσο και το χρόνο.

Το [Jakk10] ασχολείται με τον εντοπισμό μη φυσικών τιμών στην οικιακή κατανάλωση ρεύματος. Για το σκοπό αυτό καταγράφονται δεδομένα από μια οικία για χρονική περίοδο τριών μηνών. Η επιλογή για τον τρόπο καταγραφής είναι σε διαστήματα της μίας ώρας. Εκτός όμως από τα πραγματικά δεδομένα χρησιμοποιούνται και τεχνητά στα οποία έχουν «εγχυθεί» μη φυσιολογικές τιμές και καλύπτουν περίοδο 12 μηνών. Η διαδικασία δημιουργίας δεδομένων στα οποία γνωρίζουμε εκ των προτέρων την μορφολογία τους αποτελεί πρακτική που βοηθάει στην αξιολόγηση του συστήματος το οποίο έχουμε δημιουργήσει.

Στο [Kuma17] παρουσιάζεται μια προσπάθεια για εφαρμογή τεχνικών ML σε συσκευές περιορισμένων υπολογιστικών δυνατοτήτων, όπως αυτές που χρησιμοποιούνται στις εφαρμογές του IoT. Για την δοκιμή αυτών οι συγγραφείς χρησιμοποιούν σύνολα δεδομένων που βρίσκονται ελεύθερα στο Διαδίκτυο όπως τα Chars4K [DeCa09], CIFAR10 [Kriz09], MNIST [LeCu10], WARD [Yang] και άλλα. Η προοπτική και η δυνατότητα τα δεδομένα να επεξεργάζονται απευθείας στην πηγή από την οποία παράγονται είναι πολλές φορές ζωτικής σημασίας για αυτά. Σε εφαρμογές όπως αυτές της παρακολούθησης ασθενών το να μην μεταφέρονται πληροφορίες στο δίκτυο βοηθάει στη διασφάλιση των προσωπικών δεδομένων του ατόμου που παρακολουθείται από την εφαρμογή. Επίσης σε τέτοιες περιπτώσεις όπου οι αποφάσεις πρέπει να είναι γρήγορες, μειώνεται ο χρόνος απόκρισης ενώ είναι σημαντική και η δυνατότητα για λειτουργία χωρίς την ύπαρξη δικτύου. Σε αντίθετη περίπτωση για παράδειγμα, αν ένας ασθενής βρεθεί σε περιοχή χωρίς δυνατότητα σύνδεσης στο δίκτυο, αυτόματα η εφαρμογή βγαίνει εκτός λειτουργίας και μπορεί να παρέχει την άμεση ασφάλεια που χρειάζεται.

Στο [Ma13] επιχειρείται μια προσέγγιση για την εξαγωγή στοιχείων σχετικά με τη συμπεριφορά των χρηστών των μέσων μαζικής μεταφοράς. Τα δεδομένα προέρχονται από την πόλη του Πεκίνου όπου η πρόσβαση στα μέσα γίνεται με τη χρήση κάρτας. Οι 2 κατηγορίες βάσει των οποίων γίνεται η χρέωση (σταθερή και ανάλογα με την διανυόμενη απόσταση) καταγράφουν κατά την επικύρωση τους ξεχωριστά στοιχεία, τα οποία όμως δεν είναι πλήρη για την διεξαγόμενη έρευνα. Σε αυτό το σημείο καταδεικνύεται ένα γενικότερο πρόβλημα που συναντάται, όταν δηλαδή οι υπάρχουσες υποδομές που παρέχουν τα δεδομένα, δεν έχουν σχεδιασθεί με γνώμονα την ικανοποίηση των αναγκών που προκύπτουν από το IoT. Τέτοιες καταστάσεις δύναται να αντιμετωπισθούν με τη χρήση τεχνικών Μηχανικής Μάθησης και άλλων πηγών δεδομένων. Στην προκειμένη περίπτωση, το γεγονός ότι τα δεδομένα που καταγράφει η κάρτα για σταθερή χρέωση δεν περιλαμβάνει στοιχεία για την τοποθεσία επιβίβασης και αποβίβασης μπορεί να παρακαμφθεί με τη χρήση προφίλ ταχυτήτων καταγεγραμ-

μένα από GPS σε συνδυασμό με μαρκοβιανές αλυσίδες, δίνοντας μια πιθανοτική εκτίμηση σχετικά με τον τόπο επιβίβασης.

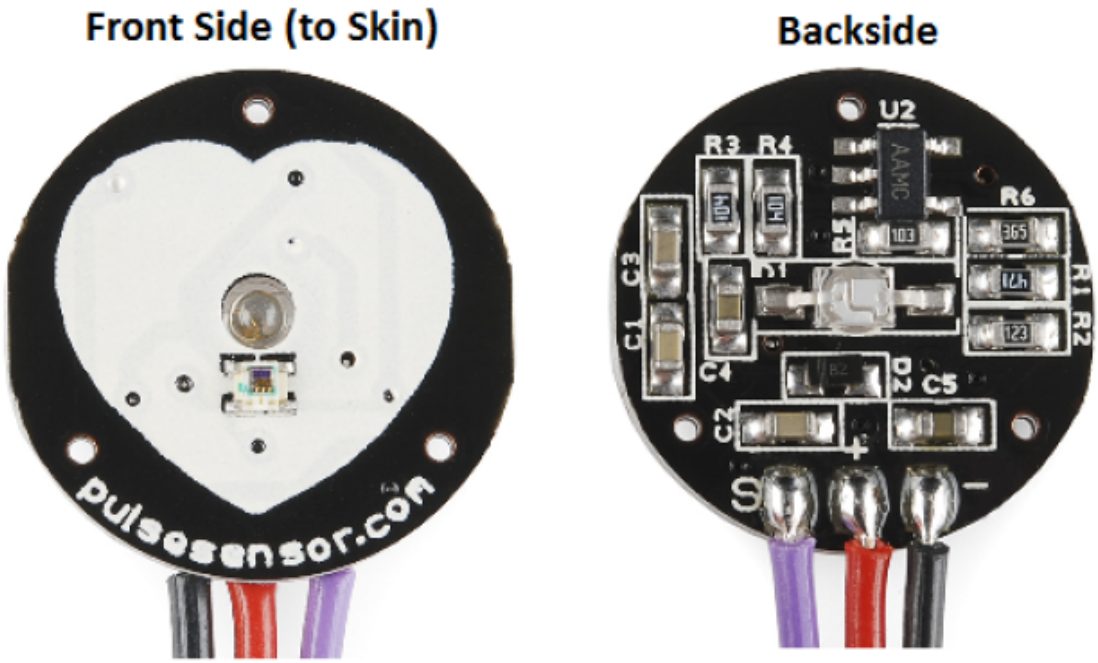
Το [Mone13] πραγματεύεται ένα γενικότερο θέμα που άπτεται του τομέα των δεδομένων. Οι αισθητήρες που χρησιμοποιούνται στις εφαρμογές είναι πιθανό να καταγράψουν τιμές εκτός των φυσιολογικών πλαισίων είτε λόγω κάποιου εκτάκτου συμβάντος είτε λόγω δικιάς τους αστοχίας στην λειτουργία. Οι συγγραφείς της εν λόγω εργασίας προτείνουν διερεύνηση του συσχετισμού τιμών μεταξύ διαφορετικών αισθητήρων αντί για την αντιμετώπιση τους ως ξεχωριστές μονάδες, με σκοπό τον εντοπισμό τέτοιων φαινομένων. Τα δεδομένα που χρησιμοποιούνται για τον έλεγχο της μεθόδου προέρχονται από καταγραφές σε «έξυπνο» σπίτι, όπου με δεδομένο ότι πρόκειται για ώρα πρωινού, γίνεται παρακολούθηση για το ποιες πόρτες είναι ανοικτές.

Τα δεδομένα στο [Ni14] προέρχονται από αισθητήρες που καταγράφουν θερμοκρασία, ατμοσφαιρική πίεση και υγρασία στο Πεκίνο. Με βάση αυτά τα δεδομένα επιχειρούν να προβλέψουν μελλοντικές τιμές των ίδιων φαινομένων. Για το σκοπό αυτό χρησιμοποιείται συνδυασμός των αλγορίθμων EEMD, SVR και PSO, με τη χρήση να μπορεί να γενικευτεί και για άλλου τύπου μετρήσεις αισθητήρων.

Στο [Pand17] τα δεδομένα αφορούν τους καρδιακούς παλμούς ατόμων. Αισθητήρες που είναι προσαρτημένοι πάνω στο άτομο καταγράφουν τους παλμούς ανά λεπτό. Για την αρχικοποίηση του συστήματος χρειάζεται μια αρχική μέτρηση για το εκάστοτε άτομο σε κατάσταση ηρεμίας και από εκεί και πέρα υπάρχει η δυνατότητα για εξαγωγή απόφασης σχετικά με την ύπαρξη στρες ή όχι. Σε αντίθεση όμως με το [Kuma17] που είδαμε παραπάνω, η επεξεργασία των δεδομένων γίνεται σε κεντρικό διακομιστή και όχι πάνω στον ίδιο τον αισθητήρα (Σχήμα 4.2). Το γεγονός αυτό κάνει το σύστημα λιγότερο ευέλικτο και αξιόπιστο.

Το [Ruta18] παρουσιάζει ένα σύστημα οδηγικής βοήθειας που καταδεικνύει την κατάσταση του οδικού δικτύου. Τα δεδομένα συλλέγονται από ενσωματωμένες πλακέτες στα οχήματα, αλλά και από το κινητό τηλέφωνο του χρήστη το οποίο πλέον είναι εξοπλισμένο με επιταχυνσιόμετρο, γυροσκόπιο και GPS. Μέσω αυτών καταγράφονται στοιχεία όπως η ταχύτητα, η αλλαγή υψομέτρου, ο φόρτος της μηχανής και η κατανάλωση καυσίμου. Επίσης κατά την εκπαίδευση του συστήματος χρειάστηκε η συμμετοχή και των οδηγών στη συλλογή στοιχείων με τη συμπλήρωση ερωτηματολογίων σχετικά με την κατάσταση του δρόμου, της κυκλοφοριακής κίνησης αλλά και το στυλ οδήγησης που συνάντησαν σε συγκεκριμένες περιπτώσεις.

Στο [Shuk15] γίνεται μελέτη για την αποδοτικότητα ορισμένων αλγορίθμων στον εντοπισμό έκτοπων τιμών πάνω σε ροές δεδομένων. Για τον έλεγχο της απόδοσης αυτών χρησιμοποιούνται τα δεδομένα του KDD CUP 99 [Tava09]. Η πρακτική της χρήσης έτοιμων δεδομένων παρουσιάζει το πλεονέκτημα της δυνατότητας για άμεση σύγκριση με τα αποτελέσματα άλλων μελετών καθώς και την απαλλαγή από την ανάγκη για συλλογή δεδομένων.



Σχήμα 4.2: Αισθητήρας καταγραφής καρδιακών παλμών [Pand17]

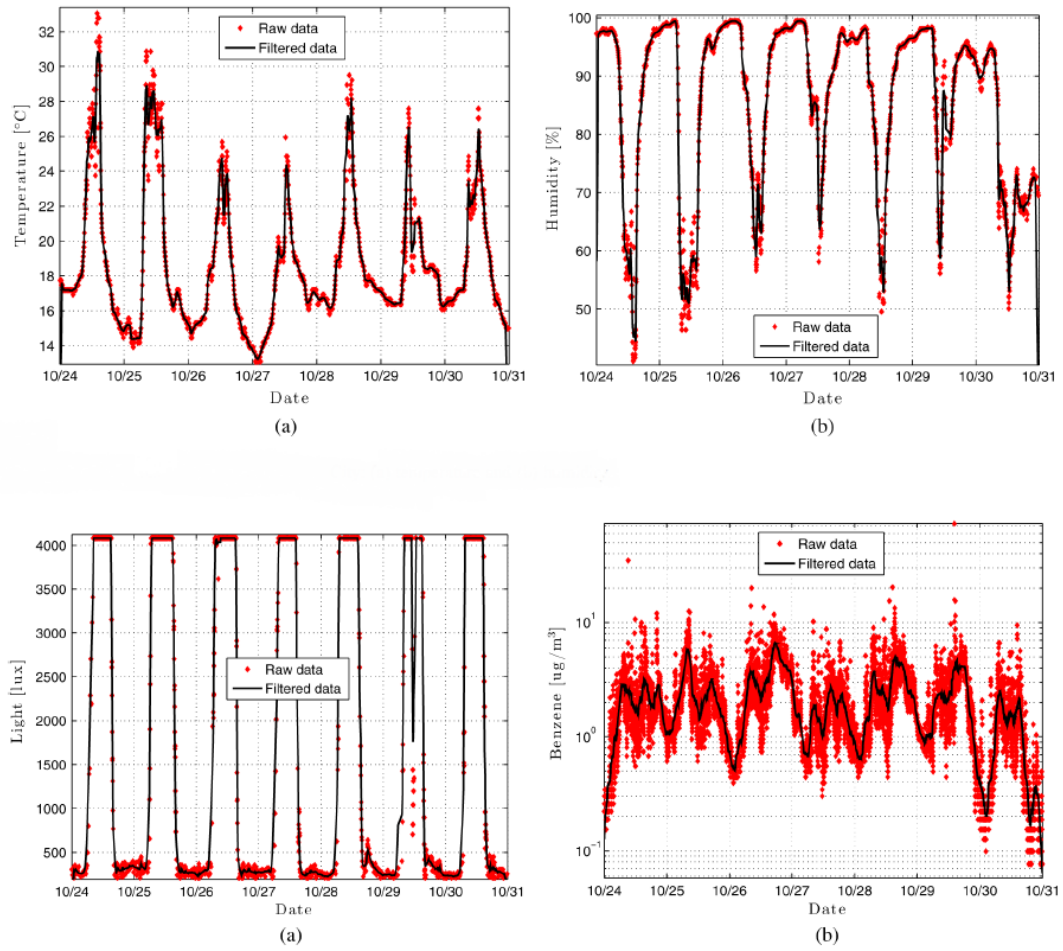
Το [Souz15] πραγματεύεται και αυτό την εύρεση μη φυσιολογικών τιμών σε ένα σύνολο δεδομένων. Τα δεδομένα που χρησιμοποιούνται προέρχονται από το European SmartSantander Project[1] και αφορούν τη θερμοκρασία και την ένταση του φωτός. Για τις ανάγκες δοκιμής του μοντέλου εκτός από τις πραγματικές τιμές προστέθηκαν και μη φυσιολογικές, οι οποίες δημιουργήθηκαν τεχνητά.

Στο [Zane12] παρουσιάζεται μια έρευνα σχετικά με την έννοια της «έξυπνης πόλης». Στα πλαίσια αυτής παραθέτονται μερικοί τομείς που θα μπορούσαν να επωφεληθούν καθώς και τα δεδομένα που καταγράφονται στην εκάστοτε περίπτωση. Για τον έλεγχο ιστορικών μνημείων αλλά και κτιρίων με έντονο ενδιαφέρον, αισθητήρες μπορούν να ελέγχουν στοιχεία της στατικότητας καθώς και περιβαλλοντικούς παράγοντες, όπως η θερμοκρασία και η υγρασία. Η καταγραφή των απορριμάτων στα σημεία που αυτά συλλέγονται μπορεί να βοηθήσει στην αποδοτικότερη αποκομιδή τους και στην εξοικονόμηση χρημάτων και ενέργειας. Ο έλεγχος της στάθμης των ατμοσφαιρικών ρύπων και του θορύβου μπορεί να καταδείξει προβλήματα που οφείλουν να καταπολεμηθούν για μια πιο βιώσιμη πόλη. Στο ίδιο πλαίσιο μπορεί να ελεγχθεί και η κυκλοφορία στους δρόμους είτε με τρόπο άμεσο, μέσω της καταγραφής των ίδιων των αυτοκινήτων, είτε έμμεσα αναγνωρίζοντας παράγοντες που σχετίζονται με αυτά όπως οι ατμοσφαιρικοί ρύποι. Παράλληλα, αναγνωρίζοντας τα επίπεδα φωτεινότητας στους δρόμους μπορεί να γίνει και αποδοτικότερη διαχείριση του φωτισμού σε αυτούς. Το τελευταίο παράδειγμα εμπίπτει στο γενικότερο πλαίσιο ενεργειακής διαχείρισης της πόλης που

[1] `http://www.smartsantander.eu/`

μπορεί να γίνει με την παρακολούθηση της συνολικής κατανάλωσης ενέργειας.

Επίσης στην εργασία παρατίθενται και τα στοιχεία από μία ανάλογη εφαρμογή στην πόλη της Πάντοβα. Αισθητήρες που είναι προσαρτημένοι στους πυλώνες φωτισμού καταγράφουν ανά τακτά χρονικά διαστήματα την θερμοκρασία, την υγρασία, τη φωτεινότητα καθώς και το επίπεδο ατμοσφαιρικών ρύπων. Στο Σχήμα 4.3 φαίνεται η απεικόνιση αυτών των δεδομένων για το διάστημα μιας εβδομάδας. Η μελέτη τους μπορεί να αποδώσει πληροφορία η οποία άλλοτε είναι προφανής και άλλοτε όχι. Για παράδειγμα το μοτίβο της αλλαγής στάθμης στα επίπεδα της φωτεινότητας ακολουθεί αυτό της εναλλαγής της μέρας με τη νύχτα. Με την ίδια λογική, αλλά όχι τόσο εμφανή τρόπο, ταυτόχρονη πτώση της θερμοκρασίας και της φωτεινότητας σε συνδυασμό με αύξηση των ρύπων μπορεί να σημαίνει ξαφνική κακοκαιρία που δημιούργησε κυκλοφοριακή συμφόρηση.

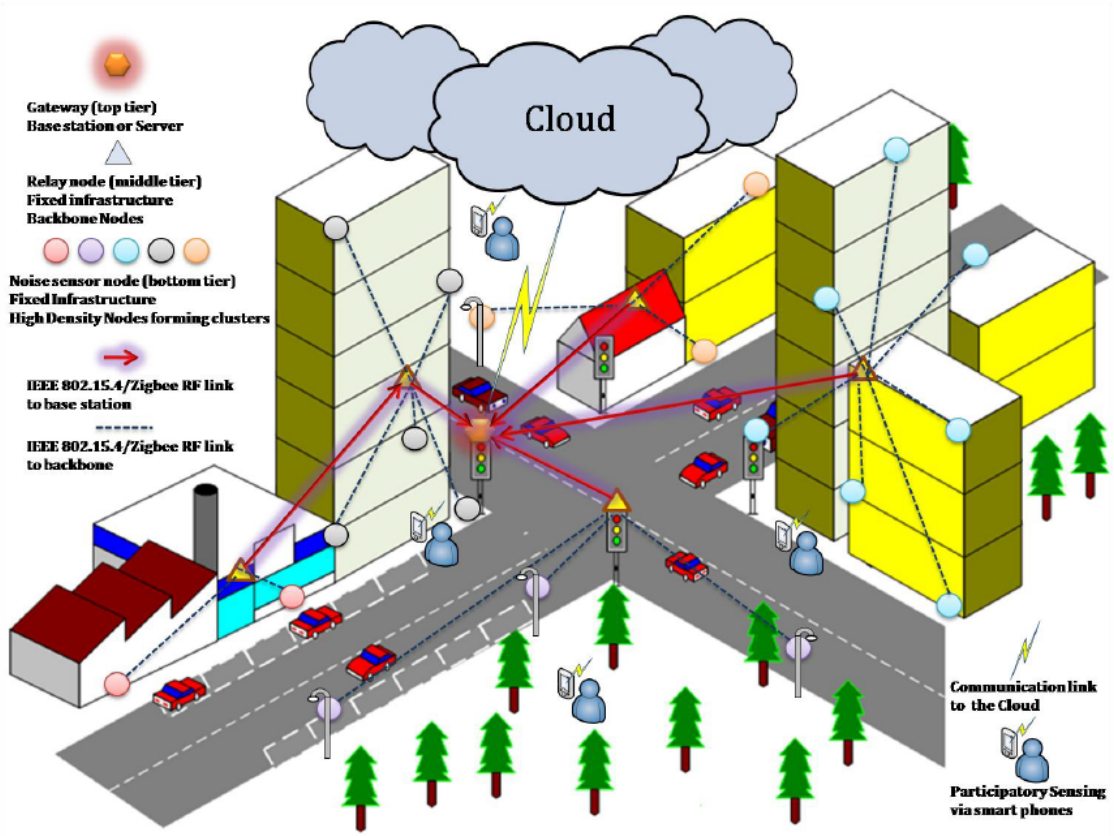


Σχήμα 4.3: Καταγραφή δεδομένων στην πόλη της Πάντοβα [Derg14]

Το [Diro18] προτείνει τρόπους για την αναγνώριση επιθέσεων στο περιβάλλον του IoT με εφαρμογή μεθόδων βαθιάς μάθησης [LeCu15]. Για το σκοπό αυτό χρησιμοποιείται η συλλογή δεδομένων NSL-KDD [Tava09], με την κάθε εγγραφή του να αποτελείται από 41 χαρακτηριστικά. Για τους σκοπούς της εργασίας χρησιμοποιήθηκε ως σύνολο δύο κλάσεων (φυ-

σιολογικό – επίθεση) αλλά και ως τεσσάρων κλάσεων (φυσιολογικό, DoS, Probe, R2L.U2R).

Το [Jin14] αποτελεί μία ακόμη εφαρμογή στα πλαίσια της «έξυπνης πόλης». Στην συγκεκριμένη περίπτωση το επίκεντρο του ενδιαφέροντος είναι η ηχορύπανση. Η ένταση του θορύβου μετριέται τόσο από αισθητήρες που βρίσκονται σε σταθερά σημεία όσο και από τους πολίτες οι οποίοι βρίσκονται σε κίνηση μέσα στην πόλη (Σχήμα 4.4). Με τον τρόπο αυτό τα δεδομένα παρουσιάζονται πιο πλήρη και με λιγότερα πιθανά χωρικά κενά. Όλες αυτές οι μετρήσεις αφού ενσωματώσουν πρώτα και το στοιχείο του χρόνου κατά τον οποίο έχουν ληφθεί αποθηκεύονται σε μια υποδομή cloud. Επίσης με τη βοήθεια του Google Maps API[2] έχει αναπτυχθεί εφαρμογή για οπτική απεικόνιση των παραπάνω, γεγονός που αυξάνει κατά πολύ τη χρηστικότητα για τον τελικό χρήστη ο οποίος έχει τα τελικά δεδομένα σε άμεσα κατανοητή μορφή.



Σχήμα 4.4: Κατανομή αισθητήρων μέσα στην πόλη [Jin14]

Στα πλαίσια της απομακρυσμένης περίθαλψης συγκαταλέγεται και η μελέτη στο [Kuma18]. Και σε αυτή την περίπτωση, αισθητήρες που είναι προσαρμοσμένοι στο σώμα του ασθενούς καταγράφουν κλινικά δεδομένα όπως η αρτηριακή πίεση και τα επίπεδα ζαχάρου στο αίμα. Κατόπιν αυτά τα δεδομένα αποθηκεύονται με κατανεμημένο τρόπο σε μια βάση Apache Hbase και από εκεί είναι διαθέσιμα για περαιτέρω επεξεργασία.

[2] `https://developers.google.com/maps/documentation/`

Παρόμοια θεματολογία και μεθοδολογία συναντάται και στο [Mano18]. Η συγκέντρωση των δεδομένων των ασθενών γίνεται μέσω αισθητήρων ενώ για την εκπαίδευση του μοντέλου χρησιμοποιούνται στοιχεία από το Cleveland Heart Disease Database (CHDD) [Detr89]. Από τις 76 ιδιότητες που περιέχονται σε αυτό το σύνολο, χρησιμοποιούνται μόνο οι 14, οι οποίες κρίνεται ότι περιέχουν την χρήσιμη πληροφορία σχετικά με την κατάσταση των ασθενών.

Στην ίδια κατηγορία ανήκει και το [Pasl15] το οποίο εστιάζει σε ασθενείς με τη νόσο Parkinson. Εκτός από τους κλινικούς δείκτες έχουμε και καταγραφή της κίνησης, η οποία βοηθάει στην παρακολούθηση της εξέλιξης της ασθένειας. Με την αποθήκευση και την επεξεργασία δεδομένων από πληθώρα ασθενών παρέχεται ταυτόχρονα υλικό ικανό να βοηθήσει στην μελέτη για την κατανόηση της φύσης της νόσου.

Στο [Hoss16] λαμβάνεται υπόψιν ένας ακόμα σημαντικός παράγοντας, η ασφάλεια των δεδομένων. Η ελεύθερη ανταλλαγή πληροφοριών πάνω στο δίκτυο μπορεί να αποτελέσει σημείο ευάλωτο σε κακόβουλες επιθέσεις. Για αυτό το λόγο στα δεδομένα ενσωματώνεται υδατογράφημα, καθιστώντας τα μη προσβάσιμα από αναρμόδια άτομα.

Εφαρμογή του IoT στον τομέα της υγεία εξετάζουν και οι συγγραφείς στο [Rahm15]. Παρότι σαν συνολική σύλληψη μοιάζει με ότι έχουμε δει έως τώρα, προτείνουν κάποιες επιπλέον διεργασίες πάνω στα δεδομένα που παράγονται στους αισθητήρες, βελτιώνοντας τη λειτουργία του συστήματος. Η πρώτη είναι η συμπίεση των δεδομένων έτσι ώστε να μειώνεται ο φόρτος της κίνησης πάνω στο δίκτυο. Η δεύτερη είναι η μετατροπή των δεδομένων σε τέτοια μορφή ώστε να μπορούν να χρησιμοποιηθούν με εύκολο τρόπο τόσο από ανθρώπους ή υπολογιστές για τη λήψη αποφάσεων. Τέλος, προτείνεται η εφαρμογή φίλτρων για την αφαίρεση πιθανού θορύβου από τα σήματα τα οποία καταγράφονται.

Το [Alam13] ελέγχει τη δυνατότητα για τη λήψη απόφασης σχετικά με το αν μια εφαρμογή σε λειτουργικό Android περιέχει κακόβουλο λογισμικό. Οι συγγραφείς χρησιμοποιούν το Amos dataset που περιέχει στοιχεία τυχαίας αλληλεπίδρασης με APK αρχεία, καθώς αυτά «τρέχουν» σε εξομοιωτή Android. Τα δεδομένα συλλέγονται σε διαστήματα των 5 δευτερολέπτων και περιέχουν στοιχεία όπως η χρήση του επεξεργαστή και της μνήμης, η θερμοκρασία και η κίνηση στο δίκτυο. Επίσης καθώς το αρχικό dataset περιείχε συντριπτικά περισσότερες εγγραφές από κακόβουλο λογισμικό, κρίθηκε απαραίτητη η εξισορρόπησή του με την εισαγωγή τεχνητά δημιουργημένων εγγραφών για μη κακόβουλο λογισμικό.

Οι συγγραφείς στο [Dong11] παρουσιάζουν μία μελέτη σχετικά με το συσχετισμό μεταξύ διαφορετικού τύπου δεδομένων που προέρχονται από αισθητήρες. Ο έλεγχος αυτός μπορεί να λαμβάνει ως βάση του και στοιχεία που προέρχονται από διαφορετικές χρονικές στιγμές. Για την δοκιμή της αποτελεσματικότητας της μεθόδου χρησιμοποιείται η συλλογή δεδομένων από το Intel Research Berkeley Lab[3]. Σε αυτό περιέχονται δεδομένα από 54 διαφορετικούς αισθητήρες, καθένας από τους οποίους καταγράφει τη θερμοκρασία, την υγρασία και την

[3] `http://db.csail.mit.edu/labdata/labdata.html`

ένταση του φωτός.

Το [Ma13] ασχολείται με την διαχείριση της οδικής κυκλοφορίας. Για το σκοπό αυτό επιστρατεύονται 4.000 ταξί τα οποία είναι εφοδιασμένα με συσκευές GPS οι οποίες καταγράφουν την χρονική στιγμή και την ακριβή τοποθεσία που βρίσκεται το όχημα. Με βάση αυτές τις καταγραφές μπορεί να υπολογιστεί η ταχύτητα του οχήματος και κατ' επέκταση το αν υπάρχει κυκλοφοριακή συμφόρηση στο συγκεκριμένο δρόμο. Αν για κάποιο σημείο δεν υπάρχουν επαρκή δεδομένα για μια δεδομένη χρονική στιγμή τότε χρησιμοποιούνται τιμές που έχουν παρουσιαστεί κατά το παρελθόν.

Τέλος, το [Uzel15] αποτελεί μελέτη σχετικά με το κατά πόσο η προσοχή των μαθητών επηρεάζεται από τις περιβαλλοντικές συνθήκες κατά τη διάρκεια διαλέξεων. Αισθητήρες καταγράφουν την θερμοκρασία, την υγρασία, την πίεση του αέρα, τη συγκέντρωση διοξειδίου του άνθρακα και την έντασή του θορύβου, με τα 2 πρώτα να αντιμετωπίζονται ως μία οντότητα για τις ανάγκες της ανάλυσης. Στη συλλογή των δεδομένων συμμετέχουν και οι ίδιοι οι μαθητές οι οποίοι ανά τακτά χρονικά διαστήματα χαρακτηρίζουν τα επίπεδα συγκέντρωσής τους.

4.2 Μετρικές

Στα προηγούμενα κεφάλαια παρουσιάσαμε εφαρμογές καθώς και τους αλγορίθμους και τις μεθόδους που χρησιμοποιούν έτσι ώστε με βάση παρελθοντικές τιμές εισόδου και εξόδου ενός συστήματος να μπορούμε να προβλέψουμε τη μελλοντική του συμπεριφορά. Η δημιουργία ενός μοντέλου όμως δεν θα είχε καμία ουσιαστική αξία αν δεν υπήρχε ένα τρόπος να ορίζουμε πόσο επιτυχημένο είναι και κατ' επέκταση πόσο μπορούμε να το εμπιστευτούμε για τις προβλέψεις μας. Για το σκοπό αυτό χρησιμοποιούνται οι μετρικές απόδοσης οι οποίες μας παρέχουν πληροφορίες σχετικά με το πόσο επιτυχημένοι είναι οι αλγόριθμοι που χρησιμοποιούμε. Η επιλογή της κατάλληλης μετρικής απόδοσης είναι σε άμεση εξάρτηση με τη φύση της εκάστοτε εφαρμογής και των δεδομένων που εμπλέκονται σε αυτή. Στο παρόν κεφάλαιο επιχειρούμε μια συνοπτική παρουσίαση των μετρικών απόδοσης που συναντήσαμε να χρησιμοποιούνται στη βιβλιογραφία, καθώς και τους λόγους για τους όποιους χρησιμοποιήθηκαν από την αντίστοιχη εφαρμογή.

4.2.1 Μέσο Απόλυτο Ποσοστιαίο Σφάλμα

Το *μέσο απόλυτο ποσοστιαίο σφάλμα* (Mean Absolute Percentage Error ή MAPE) δίνεται στην Εξίσωση 4.1

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| (\%) \tag{4.1}$$

όπου A_t είναι η πραγματική τιμή, F_t είναι η τιμή της πρόβλεψης και n ο συνολικός αριθμός των παρατηρήσεων. Πρόκειται για το μέσο όρο της ποσοστιαίας απόκλισης της προβλεπόμενης από την πραγματική τιμή. Στα θετικά της συγκαταλέγονται ότι είναι απλή στην εφαρμογή και στην κατανόηση. Στον αντίποδα, το βασικό της μειονέκτημα και πρόβλημα είναι ότι δεν μπορεί να χειριστεί περιπτώσεις που περιέχουν μηδενικές τιμές λόγω του ότι αυτές θα πρέπει να χρησιμοποιηθούν ως παρανομαστής σε διαίρεση. Επίσης για πολύ χαμηλές τιμές πρόβλεψης η τιμή της δεν μπορεί να ξεπεράσει το 100% ενώ για πολύ υψηλές δεν υπάρχει κάποιο άνω φράγμα. Η συγκεκριμένη μετρική χρησιμοποιείται στις εργασίες [Derg14] και [Ni14].

4.2.2 Μέσο Τετραγωνικό Σφάλμα

Το μέσο τετραγωνικό σφάλμα (Mean Squared Error ή MSE) δίνεται στην Εξίσωση 4.2

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2 \tag{4.2}$$

όπου και πάλι A_t είναι η πραγματική τιμή ενώ F_t είναι η τιμή της πρόβλεψης. Όπως προκύπτει και από την ονομασία της, πρόκειται για τη μέση τιμή των τετραγώνων των σφαλμάτων που έχουν προκύψει από τις προβλέψεις. Η τιμή της είναι πάντα θετική και όσο πιο κοντά βρίσκεται στο μηδέν, τόσο καλύτερη είναι η επίδοση του μοντέλου το οποίο ελέγχει. Στα ενδιαφέροντα στοιχεία της συγκαταλέγεται και το γεγονός ότι καθώς τα σφάλματα υψώνονται στο τετράγωνο, τα μεγαλύτερα από αυτά έχουν αυξημένη συνεισφορά στην «ποινή» για το τελικό άθροισμα σε σχέση με τα μικρότερα. Και αυτή έχει εύκολη εφαρμογή ενώ προσφέρει μια εποπτική και εύκολα κατανοήσιμη εικόνα για την απόδοση. Την εν λόγω μετρική την συναντάμε στις εργασίες [Ni14] και [Mano18].

4.2.3 Ρίζα του Μέσου Τετραγωνικού Σφάλματος

Η *ρίζα του μέσου τετραγωνικού σφάλματος* (Root Mean Squared Error ή RMSE) ισούται με την τετραγωνική ρίζα του MSE ($RMSE = \sqrt{MSE}$) και έχει παρόμοια χρήση με αυτή. Η τιμή της δίνεται στην Εξίσωση 4.3.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2} \tag{4.3}$$

Χρήση της γίνεται στα [Derg14], [Alam13] και [Mano18].

4.2.4 Άθροισμα Τετραγωνικών Σφαλμάτων

Το *άθροισμα τετραγωνικών σφαλμάτων* (Sum of Squared Errors ή SSE) δίνεται στην Εξίσωση 4.4

$$SSE = \sum_{t=1}^{n}(A_t - F_t)^2 \tag{4.4}$$

όπου A_t είναι η πραγματική τιμή και F_t είναι η τιμή της πρόβλεψης. Επίσης ισχύει ότι $SSE = n * MSE$, όπου MSE είναι η μετρική της Εξίσωσης 4.2. Χρήση της παρατηρούμε στο [Mano18].

4.2.5 Πίνακας Σύγχυσης

Ο *πίνακας σύγχυσης* (Confusion Matrix) χρησιμοποιείται σε προβλήματα κατηγοριοποίησης. Πρόκειται για ένα πίνακα ο οποίος στις γραμμές του καταγράφει τις προβλεπόμενες κλάσεις και στις στήλες τις πραγματικές κλάσεις (ή το ανάποδο). Όπως φαίνεται στο Σχήμα 4.5 που αποτελεί παράδειγμα πίνακα σύγχυσης για σύγκριση διαφορετικών αλγορίθμων, παρέχεται οπτική απεικόνιση των αποτελεσμάτων της κατηγοριοποίησης. Με πράσινο χρώμα φαίνονται οι σωστές προβλέψεις ενώ είναι εύκολο να διαπιστώσουμε και αν ο αλγόριθμος μας δυσκολεύεται στο να ξεχωρίσει κάποιες κατηγορίες μεταξύ τους. Την τεχνική αυτή τη συναντάμε στα [Alam16] και [Pand17]

4.2.6 Ακρίβεια

Η *ακρίβεια* (Accuracy) αποτελεί τη διαφορά μεταξύ της προβλεπόμενης από την πραγματική τιμή. Όσο πιο μικρή είναι αυτή η διαφορά, τόσο καλύτερο το αποτέλεσμα. Στην περίπτωση προβλημάτων κατηγοριοποίησης δίνεται από την Εξίσωση 4.5

$$Accuracy = \frac{Numberofcorrectpredictions}{Totalnumberofpredictions} \tag{4.5}$$

Η συγκεκριμένη μετρική όμως εστιάζει απολύτως στο τελικό αποτέλεσμα χωρίς να λαμβάνει υπόψιν καθόλου την κατανομή των κλάσεων που εμπλέκονται στην πρόβλεψη. Για παράδειγμα αν έχουμε 2 κλάσεις A και B σε ένα δείγμα με την πρώτη να εμφανίζεται σε ποσοστό 99% και τη δεύτερη σε ποσοστό 1%, θα μπορούσαμε απλά να προβλέπουμε μόνιμα την κλάση A πετυχαίνοντας πολύ υψηλό δείκτη ακρίβειας. Παρά την φαινομενικά υψηλή επίδοση, δεν έχουμε καμία δυνατότητα για να προβλέψουμε την εμφάνιση της κλάσης B. Υπάρχουν όμως περιπτώσεις, όπως σε ιατρικές εφαρμογές, όπου κάποια ασθένεια μπορεί να είναι εξαιρετικά σπάνια, όμως η δυνατότητα σωστή πρόβλεψη της να είναι πολύ σημαντική. Για αυτό το λόγο ο συγκεκριμένο δείκτης είναι χρήσιμος μόνο όταν εξετάζουμε σύνολα με ισομερή κατανομή κλάσεων. Χρησιμοποιείται στα [Diro18], [Hoss16] και [Uzel15].

Actual/Predicted	Sitting	Sitting Down	Standing	Standing Up	Walking
Sitting	50594	3	1	33	1
Sitting Down	12	11523	139	103	50
Standing	2	16	47127	82	143
Standing Up	48	260	267	11806	34
Walking		106	979	85	42220

a

Actual/Predicted	Sitting	Sitting Down	Standing	Standing Up	Walking
Sitting	46023	457	3885	258	8
Sitting Down	1078	6838	3084	174	653
Standing	306	614	43852	146	2452
Standing Up	1099	2733	5117	1658	1808
Walking	588	2127	8820	2623	29232

b

Actual/Predicted	Sitting	Sitting Down	Standing	Standing Up	Walking
Sitting	25366	9	1	9	3
Sitting Down	1	5825	48	59	16
Standing	1	4	23470	44	36
Standing Up	14	106	93	5975	23
Walking	5	67	280	55	21337

c

Actual/Predicted	Sitting	Sitting Down	Standing	Standing Up	Walking
Sitting	50622			9	
Sitting Down	5	11720	18	53	31
Standing		18	47252	26	74
Standing Up	9	55	35	12264	52
Walking	1	26	73	36	43254

d

Actual/Predicted	Sitting	Sitting Down	Standing	Standing Up	Walking
Sitting	50515	35	2	77	2
Sitting Down	1215	6131	2574	1403	504
Standing		353	43305	61	3651
Standing Up	1448	2002	2789	5579	597
Walking	59	209	12254	834	30034

e

Actual/Predicted	Sitting	Sitting Down	Standing	Standing Up	Walking
Sitting	50616	1	1	13	
Sitting Down	13	11666	52	67	29
Standing		6	47253	24	87
Standing Up	24	90	66	12189	46
Walking		23	69	24	43274

f

Actual/Predicted	Sitting	Sitting Down	Standing	Standing Up	Walking
Sitting	50583	6		22	1
Sitting Down	9	11437	31	250	96
Standing		132	47096	121	237
Standing Up	39	173	74	11951	105
Walking		79	169	71	42951

g

Actual/Predicted	Sitting	Sitting Down	Standing	Standing Up	Walking
Sitting	25331	1		8	
Sitting Down		5808	1	62	10
Standing		20	23738	75	61
Standing Up	12	34	42	6038	14
Walking		8	35	11	21507

h

Fig. 2. Confusion matrix of (a) SVM; (b) KNN; (c) NB; (d) C4.5, (e) LDA; (f) C5.0; (g) ANNs and (h) DLANNs for dataset [29]

Σχήμα 4.5: Παράδειγμα πίνακα σύγχυσης [Alam16]

4.2.7 Πιστότητα

Η *πιστότητα* (Precision) δείχνει το κατά πόσο οι προβλεπόμενες τιμές βρίσκονται κοντά μεταξύ τους. Σε προβλήματα δυαδικής ταξινόμησης, η πιστότητα δίνεται από την Εξίσωση 4.6 και τον συναντάμε στο [Diro18] και στο [Uzel15]. Πρόκειται δηλαδή για το ποσοστό των σωστών προβλέψεων ότι ένα δείγμα ανήκει στην κατηγορία ανάμεσα σε όλες τις προβλέψεις (σωστές και λάθος) που υποδεικνύουν ότι ένα δείγμα ανήκει στην ίδια κατηγορία. Με δεδομένη την αναζήτηση για τα στοιχεία μιας κατηγορίας, είναι δηλαδή μια ένδειξη για το τι πιθανότητα υπάρχει αν ένα στοιχείο έχει προβλεφθεί ότι ανήκει στην κατηγορία, να ανήκει όντως σε αυτή.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4.6}$$

4.2.8 Ανάκληση και Ευαισθησία

Η *ανάκληση* (Recall) δίνεται στην Εξίσωση 4.7

$$Recall = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4.7}$$

Πρόκειται για το σύνολο των σωστών προβλέψεων ότι κάποιο δεδομένο ανήκει στην κατηγορία, διαιρούμενο από το σύνολο των δεδομένων που ανήκουν στην κατηγορία, είτε αυτά έχουν χαρακτηρισθεί σωστά ότι ανήκουν σε αυτή είτε λανθασμένα ότι δεν ανήκουν. Σε αντίθεση λοιπόν με την ακρίβεια, η ανάκληση παρουσιάζει πόσο πλήρης είναι μια αναζήτηση για τα στοιχεία κάποιας κατηγορίας, το μέτρο δηλαδή στο οποίο ο αλγόριθμος είναι ικανός να ανακαλύπτει όλα τα στοιχεία που ανήκουν σε αυτή. Στην περίπτωση της δυαδικής ταξινόμησης ονομάζεται *ευαισθησία* (Sensitivity). Τον συναντάμε στο [Uzel15] και στο [Kuma18].

4.2.9 Ειδικότητα

Η *ειδικότητα* (Specificity) δίνεται από την Εξίσωση 4.8.

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{4.8}$$

Παρουσιάζει τη βεβαιότητα με την οποία ένας αλγόριθμος μπορεί να αποφανθεί ότι κάποια περίπτωση δεν ανήκει σε μια συγκεκριμένη κλάση. Κάτι τέτοιο είναι ιδιαίτερο χρήσιμο στην ιατρική, όπως και στο [Kuma18], όπου τον βλέπουμε να χρησιμοποιείται. Αν για παράδειγμα θέλουμε να αποφασίσουμε αν κάποιο άτομο έχει μια ορισμένη ασθένεια είναι σημαντικό να γνωρίζουμε ότι ένα αρνητικό αποτέλεσμα έχει μικρή πιθανότητα να είναι λάθος. Σε αντίθετη περίπτωση, όπου προβλέπαμε δηλαδή μη ύπαρξη αλλά το άτομο είναι ασθενής, οι επιπτώσεις μπορεί να είναι ολέθριες.

4.2.10 F Score

Το F_1 score είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης και ο τύπος του δίνεται στην Εξίσωση 4.9

$$F_1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.9}$$

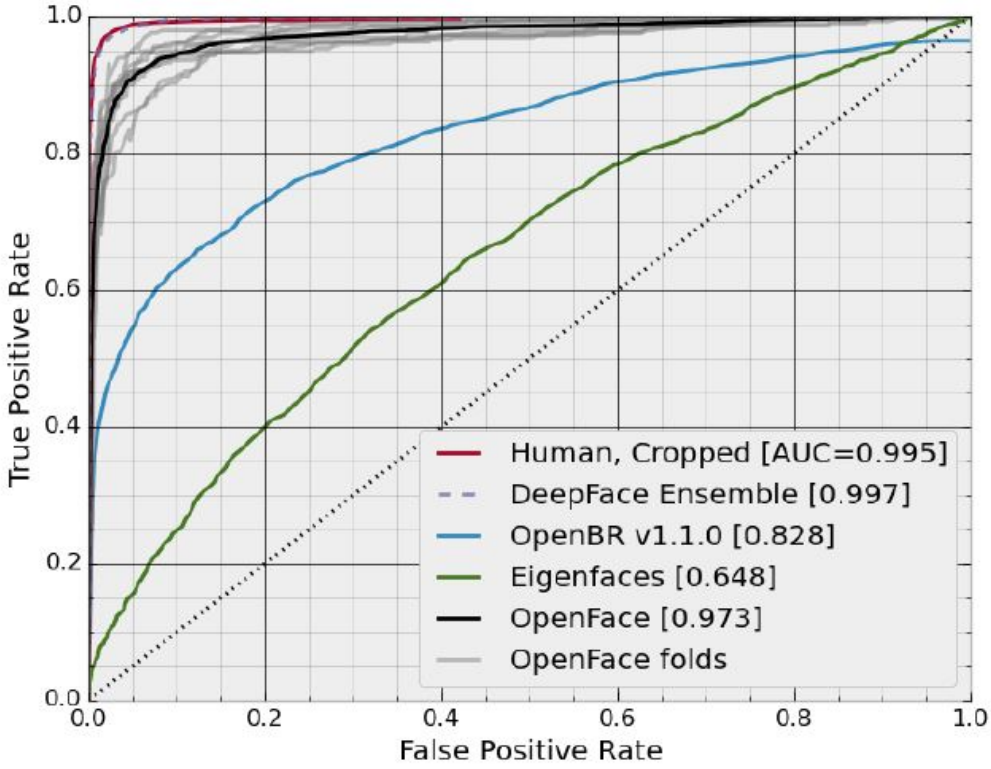
Οι τιμές του βρίσκονται στο εύρος $[0, 1]$ με τις μεγαλύτερες να καταδεικνύουν καλύτερα αποτελέσματα. Σαν δείκτης παρουσιάζει ταυτόχρονα το πόσο ακριβής είναι ένας αλγόριθμος στις προβλέψεις του αλλά και το κατά πόσο καταφέρνει να μην αγνοεί τις εμφανίσεις περιπτώσεων κάποια κλάσης. Το γεγονός όμως ότι χρησιμοποιεί εξίσου την ακρίβεια και την ανάκληση μπορεί να αποτελέσει παράγοντα αδυναμίας, καθώς αυτές, ανάλογα με το πρόβλημα, είναι πιθανό να μην έχουν την ίδια βαρύτητα στη φυσική του ερμηνεία. Το F score, στη γενική του μορφή, δίνεται από την Εξίσωση 4.10, με την οποία μπορεί να ξεπεραστεί το παραπάνω πρόβλημα.

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall} \tag{4.10}$$

Συνηθισμένες περιπτώσεις είναι οι F_2 και $F_{0.5}$ οι οποίες δίνουν περισσότερη βάση στην ανάκληση και στην ακρίβεια, αντίστοιχα. Χρήση της F_1 συναντάμε στο [Diro18], όπου αξιολογεί νευρωνικά δίκτυα τα οποία αναγνωρίζουν δικτυακές επιθέσεις.

4.2.11 Περιοχή Κάτωθεν της Καμπύλης

Η *περιοχή κάτωθεν της καμπύλης* (Area Under Curve ή AUC) είναι μια μετρική που χρησιμοποιείται σε προβλήματα δυαδικής κατηγοριοποίησης. Πρόκειται για την πιθανότητα ένας κατηγοριοποιητής να κατατάξει ένα τυχαίο θετικό δείγμα πιο πάνω από ένα τυχαίο αρνητικό δείγμα. Παράδειγμα χρήσης της βλέπουμε στο Σχήμα 4.6, όπου οι ρυθμοί *αληθώς θετικών* (true positive) και *ψευδώς θετικών* (false positive) που σχηματίζουν την καμπύλη ROC υπολογίζονται με τη χρήση οριακών τιμών στην κατανομή των δειγμάτων. Οι τιμές της βρίσκονται στο εύρος $[0, 1]$ με τις μεγαλύτερες να δείχνουν καλύτερη απόδοση του μοντέλου.



Σχήμα 4.6: Παράδειγμα περιοχής κάτωθεν της καμπύλης [Amos16a]

Κεφάλαιο 5

Συμπεράσματα και Μελλοντικές Κατευθύνσεις

5.1 Συμπεράσματα

Ο όρος IoT ξεκίνησε να υπάρχει σε μια εποχή που ακόμη και το Διαδίκτυο δεν ήταν πλήρως διαδεδομένο και προσβάσιμο από το μεγαλύτερο μέρος του πληθυσμού. Ακόμα και αν υπήρχε σαν ιδέα, το γεγονός ότι δεν υπήρχαν τα τεχνολογικά μέσα που θα επιτρέψουν την εφαρμογή του σε μεγάλη κλίμακα το έκανε να φαντάζει ως κάτι μακρινό. Τα τελευταία χρόνια με την εξάπλωση των «έξυπνων» συσκευών αλλά και την εξέλιξη των ενσωματωμένων συστημάτων και των δικτύων επικοινωνίας αυτό άλλαξε. Το IoT δεν αποτελεί πια το μέλλον αλλά είναι το παρόν με το ενδιαφέρον γύρω από αυτό να γίνεται ολοένα και μεγαλύτερο. Από το 2010 και μετά έχουν αυξηθεί οι ερευνητικές εργασίες, μέρος των οποίων παρουσιάσαμε στην παρούσα εργασία με σκοπό να εξάγουμε χρήσιμα συμπεράσματα σχετικά με την πορεία και τις τάσεις σε αυτές.

Όσον αφορά τα πεδία εφαρμογής παρατηρούμε ότι το μεγαλύτερο ενδιαφέρον παρουσιάζει η «έξυπνη πόλη». Θα μπορούσαμε να πούμε ότι πρόκειται για την πρώτη μεγάλης κλίμακας εφαρμογή που μπορεί να βρει πρακτική υλοποίηση. Σε αυτό το γεγονός συμβάλει το ότι μιλάμε για μια δομή (πόλη) με περιορισμένο χωρικό εύρος αλλά και ευχέρεια στο σχεδιασμό των υποδομών που θα συνδράμουν στην επίτευξη κάποιου στόχου. Πόλεις όπως η Μελβούρνη και το Σανταντέρ υλοποιούν ήδη σχετικές εφαρμογές με σκοπό τη βελτίωση συνθηκών ζωής των πολιτών τους.

Άλλος κλάδος ο οποίος συγκεντρώνει έντονο ερευνητικό ενδιαφέρον είναι αυτό της υγείας. Ο αυτοματοποιημένος τρόπος της διάγνωσης και αντιμετώπισης σοβαρών ασθενειών μέσω συσκευών που δεν απαιτεί τη φυσική παρουσία του ασθενή σε νοσοκομείο αποτελεί σοβαρό κίνητρο για την ανάπτυξη προς την κατεύθυνση αυτή. Σε αντίθεση όμως με την περίπτωση τον πόλεων, εδώ οι μελέτες περιορίζονται κατά κύριο λόγο σε θεωρητικό επίπεδο και σε προτάσεις για συστήματα που θα μπορούσαν να υλοποιηθούν στην πράξη.

Άλλοι δύο κλάδοι που παρουσιάζουν σημαντικό ενδιαφέρον αλλά και πρακτικό έργο είναι αυτός της διαχείρισης της κυκλοφοριακής κίνησης και το «έξυπνο σπίτι». Επίσης παρατηρούμε ότι λόγω του μεγάλου εύρους δυνατοτήτων που προσφέρει το IoT μπορούν να

προκύψουν εφαρμογές οι οποίες δεν εντάσσονται απαραίτητα σε κάποια γενικότερη κατηγορία, όπως ο έλεγχος για τη διασφάλιση της ποιότητας των τροφίμων.

Μία από τις προκλήσεις που παρουσιάζονται στην εφαρμογή όσων παρουσιάσαμε παραπάνω είναι η αρχιτεκτονική πάνω στην οποία θα στηθεί το εκάστοτε σύστημα.Εξαιτίας του πλήθους των συσκευών που εμπλέκονται σε αυτό αλλά και τις περιορισμένες υπολογιστικές δυνατότητες που έχουν, πρέπει να υπάρξει ειδική μέριμνα ώστε να είναι εφικτές η καλύτερη δυνατή επικοινωνία μεταξύ τους αλλά και η αποδοτική επεξεργασία των δεδομένων τα οποία παράγονται. Για αυτό το σκοπό έχουν προταθεί και παρουσιασθεί διάφορες προσεγγίσεις οι βασικοί άξονες των οποίων έχουν να κάνουν με την ύπαρξη ενδιάμεσων σταθμών που θα συλλέγουν τα δεδομένα από τις συσκευές στις οποίες παράγονται, με σκοπό την αποστολή τους σε κάποιο κεντρικό διακομιστή για επεξεργασία ή τη δυνατότητα των συσκευών να μπορούν να επικοινωνούν άμεσα μεταξύ τους και να εκτελούν υπολογισμούς μειώνοντας έτσι τη συμφόρηση του δικτύου.

Επίσης σημαντικό ρόλο και πρόκληση στην υλοποίηση του IoT αποτελούν τα δεδομένα που αυτό διαχειρίζεται. Το γεγονός ότι αυτά προέρχονται από πληθώρα διαφορετικών πηγών εισάγει στο σύστημα σημαντικά προβλήματα ως προς την ενιαία αντιμετώπισή τους. Αποτελεί ανοικτό πρόβλημα η συστηματική αντιμετώπιση του φαινομένου αυτού και η μετατροπή των δεδομένων σε μορφή που να εμπεριέχουν όσο το δυνατό περισσότερη χρήσιμη πληροφορία αλλά και να είναι «συμβατά» μεταξύ τους.

Άλλο ένα θέμα που προκύπτει σχετικά με τα δεδομένα είναι ο μεγάλος όγκος τους σε σχέση με κλασσικές εφαρμογές. Το γεγονός αυτό εισάγει δυσκολίες στον τρόπο μετάδοσης, αποθήκευσης και επεξεργασίας τους. Για αυτό το λόγο παρατηρούμε ότι αρκετές μελέτες προτείνουν τη χρήση υποδομών για κατανεμημένα συστήματα αποθήκευσης και επεξεργασίας όπως το Apache Hadoop [Vavi13].

Σημείο διαφοροποίησης με τις κλασσικές στατικές εφαρμογές είναι ότι σε πολλές περιπτώσεις τα δεδομένα καταφθάνουν από ζωντανές ροές. Οι ροές δεδομένων φέρουν ιδιαιτερότητες οι οποίες απαιτούν ειδική διαχείριση για την επιτυχή ανάλυσή τους και για αυτό αποτελούν και ανοιχτό σημείο για έρευνα. Ο ρυθμός με τον οποίο καταφθάνουν τα δεδομένα δεν είναι σταθερός ενώ και η συνεχή παραγωγή τους μπορεί να χαρακτηρίσει της ροές ως άπειρες από την πλευρά του χρήστη που πρέπει να μπορεί να αποφασίσει πότε έχει αποκτήσει ικανοποιητική ποσότητα πληροφορίας ώστε να μπορεί να εξάγει συμπεράσματα. Για τον ίδιο λόγο ορισμένες φορές η επεξεργασία του dataset πρέπει να γίνει υποχρεωτικά με μόνο ένα πέρασμα-διάβασμα των δεδομένων καθώς αυτά στην πορεία παύουν να είναι διαθέσιμα. Ένας ακόμα παράγοντας που μπορεί να επηρεάσει το σύστημα είναι η περιοδικότητα στην ισχύ και στην εγκυρότητα των δεδομένων. Κάποια από αυτά μπορεί μετά από ένα χρονικό διάστημα να μην είναι αντιπροσωπευτικά για το μοντέλο που θέλουμε να δημιουργήσουμε, ενώ είναι πιθανό το νόημά τους να είναι άμεσα συνδεδεμένο με το περιβάλλον τους το οποίο

όμως και αυτό ενδέχεται να μεταβάλλεται δυναμικά. Επίσης τα δεδομένα προς επεξεργασία μπορεί να προέρχονται από περισσότερες από μία ροές και να φτάνουν στο σύστημα με ασύγχρονο τρόπο. Έτσι πέρα από το θέμα της ετερογένειας που έχει ήδη αναφερθεί, εισάγεται και η ανάγκη για συγχρονισμό των δεδομένων. Από την παρουσίαση του ερευνητικού έργου που κάναμε στα πλαίσια της παρούσας εργασίας, προκύπτει ότι τα ζητήματα που αναφέρουμε σε αυτή την παράγραφο προσελκύουν αρκετά έντονο ενδιαφέρον από μερίδα ερευνητών. Μάλιστα το συγκεκριμένο κομμάτι ξεπερνάει τα πλαίσια του IoT και αποκτάει αξία και ως αυτόνομη οντότητα, καθώς σε μια εποχή που οι ρυθμοί με τους οποίους παράγεται η πληροφορία συνεχώς αυξάνονται, η μελέτη για την ανάλυση και επεξεργασία των ροών δεδομένων είναι ιδιαίτερα σημαντική.

Η μηχανική μάθηση αποτελεί θεμέλιο λίθο τόσο για τη δημιουργία όσο και την εξέλιξη του IoT. Σαν επιστημονικός κλάδος που προϋπάρχει και ασχολείται με τη μετατροπή της πληροφορίας σε χρήσιμη γνώση, του παρέχει τα απαραίτητα εργαλεία έτσι ώστε να μπορέσει να αξιοποιήσει τα δεδομένα που παράγονται μέσα στο δίκτυο του. Το IoT όμως εκτός από το να χρησιμοποιεί την μηχανική μάθηση συμβάλλει και αυτό σαν αφορμή για περαιτέρω έρευνα και εξέλιξη του κλάδου σε νέα επίπεδα. Έτσι βλέπουμε να υπάρχει ενδιαφέρον για την υλοποίηση ήδη γνωστών αλγορίθμων σε κατανεμημένο περιβάλλον καθώς και την προσαρμογή γνωστών τεχνικών έτσι ώστε να μπορούν να ανταποκρίνονται στις απαιτήσεις των ροών δεδομένων που αναφέραμε προηγουμένως.

Έχοντας πλέον μια ολοκληρωμένη εικόνα του IoT παρατηρούμε ότι επί της ουσίας αποτελεί ένα συνδυασμό πληθώρας τεχνολογικών και επιστημονικών κλάδων μέσω των οποίων επιτυγχάνεται το επιθυμητό αποτέλεσμα. Η Μηχανική Μάθηση δεν θα χρησίμευε σε κάτι αν δεν υπήρχε η εξέλιξη των αισθητήρων, οι οποίοι τροφοδοτούν με δεδομένα, όπως αντίστοιχα και οι αισθητήρες θα ήταν άνευ χρησιμότητας αν δεν υπήρχε η μηχανική μάθηση για να αξιοποιήσει τα δεδομένα που αυτοί παράγουν. Το πλέον ενδιαφέρον στη συγκεκριμένη υπόθεση είναι ότι η ανάπτυξη και εξέλιξη του ενός δημιουργεί νέα πεδία και προοπτικές ενδιαφέροντος για τα άλλα, ρίχνοντας νερό στο μύλο της έρευνας. Αν και η μεγάλη αλληλεξάρτηση μεταξύ τους δυσκολεύει τις προβλέψεις, στην επόμενη και τελευταία υποενότητα θα παρουσιάσουμε μια σειρά από πιθανές μελλοντικές κατευθύνσεις στην εξέλιξη του IoT.

5.2 Μελλοντικές Κατευθύνσεις

Οι προβλέψεις σχετικά με το μέλλον στον τομέα της τεχνολογίας είναι πάντα δύσκολες. Οι εξελίξεις είναι ραγδαίες και οι συσχετισμοί μπορούν να αλλάξουν μέσα σε ελάχιστο χρονικό διάστημα. Αυτό που παρατηρούμε με το IoT είναι ότι έχει περάσει από την πρώτη επαναστατική εποχή του και πλέον οδεύει προς τη φάση της ωριμότητας. Οι πρώτες ιδέες έχουν περάσει από το πεδίο της θεωρίας σε αυτό της υλοποίησης και πλέον αναμένεται με-

γάλο μέρος τους ερευνητικού ενδιαφέροντος να εστιάσει στην ανάπτυξη και εξέλιξη αυτών.

Η «έξυπνη πόλη» έχει παρουσιάσει ήδη σημαντικές εφαρμογές οι οποίες όμως περιορίζονται σε μικρό χωρικό πλαίσιο. Το επόμενο βήμα είναι η βελτίωση τον ήδη υπαρχόντων παροχών προς του πολίτες αλλά και η επέκταση της ιδέας σε μεγαλύτερη κλίμακα (για παράδειγμα χώρας) όπου αυτό είναι εφικτό και έχει νόημα. Από την άλλη στον τομέα της υγείας αναμένουμε να δούμε υλοποιήσεις που θα παρέχουν ένα πλήρες και αξιόπιστο πακέτο στον χρήστη, χωρίς τα μέσα που θα χρησιμοποιούνται να τον εμποδίζουν στην καθημερινότητά του.

Στο τεχνολογικό κομμάτι και στο θέμα των υποδομών που υποστηρίζουν το IoT υπάρχει συνεχής εξέλιξη. Ο αριθμός των συσκευών που έχει δυνατότητα πρόσβασης στο Διαδίκτυο αυξάνεται με σταθερό ρυθμό ενώ αυτές αποκτούν ολοένα και περισσότερη υπολογιστική δύναμη, όντας παράλληλα και πιο προσιτές οικονομικά. Βέβαια καθώς το υλικό δεν είναι ο μοναδικός παράγοντας που παίζει ρόλο, δεν μπορούμε να αναμένουμε ότι η όποια αύξηση στην «ποιότητα» θα επηρεάσει με γραμμικό τρόπο και την συνολική ποιότητα των συστημάτων.

Έως τώρα οι υλοποιήσεις που είχαν ουσιαστικά αποτελέσματα έχουν από πίσω τους κατά κύριο λόγο μεγάλους οργανισμούς, καθώς μόνο τέτοιοι μπορούν να διαθέτουν τα ανάλογα μέσα. Μικρότερης κλίμακας προσπάθειες, αν και σε αρκετές περιπτώσεις έχουν έντονο ερευνητικό ενδιαφέρον, δεν μπορούν να παράξουν ανάλογο αντίκτυπο ως προς τη χρησιμότητα των πρακτικών εφαρμογών τους, οι οποίες περιορίζονται σε στενά πλαίσια ως προς τη λειτουργικότητα τους. Κάτι που μπορεί να αλλάξει την κατάσταση αυτή είναι η εξέλιξη των εργαλείων ανάπτυξης αλλά και η προτυποποίηση ορισμένων εκ των συνηθέστερων διεργασιών. Κάτι ανάλογο είχε παρατηρηθεί και στον χώρο του Διαδικτύου, που από την στιγμή που δημιουργήθηκαν τα εργαλεία που απλουστεύουν την παραγωγή, «εκτοξεύθηκε» και η χρηστικότητα του.

Ο τομέας της αρχιτεκτονικής είναι από αυτούς που δεν αναμένονται ιδιαίτερες εξελίξεις, τουλάχιστον όχι στο άμεσο μέλλον. Τα κατανεμημένα συστήματα είναι ένας κλάδος που έχει σημειώσει μεγάλη πρόοδο τα τελευταία χρόνια και όπως βλέπουμε είναι ικανός να προσφέρει τις απαιτούμενες δυνατότητες αποθήκευσης και επεξεργασίας όπου αυτό απαιτείται.

Από την άλλη το κομμάτι των δεδομένων αναμένεται να προσελκύσει το ενδιαφέρον των ερευνητών τα επόμενα χρόνια καθώς παρουσιάζει το μεγαλύτερο περιθώριο για βελτίωση και καινοτομίες. Αυτή τη στιγμή η ετερογένεια των δεδομένων είναι ένας από τους βασικούς λόγους που οι εφαρμογές περιορίζονται συνήθως σε μόνο μία θεματική ενότητα και σε όχι πολύ μεγάλο εύρος. Η εύρεση τρόπου τέτοιου ώστε τα δεδομένα που παράγονται από τους αισθητήρες να μετατρέπονται με συστηματικό και αξιόπιστο τρόπο σε μορφή που θα αναπαριστά το γενικότερο πλαίσιο στο οποίο αναφέρονται και θα είναι κατανοητά από τους ανθρώπους και τις άλλες μηχανές, θα ανοίξει τεράστιους ορίζοντες για την εξέλιξη του

IoT.

Τέλος, η Μηχανική Μάθηση βρίσκεται στο στάδιο ωριμότητάς της. Αυτό σημαίνει ότι δεν αναμένουμε να δούμε για παράδειγμα την εμφάνιση κάποιου αλγορίθμου που θα φέρει δραστικές αλλαγές στον τρόπο που βλέπουμε και αντιλαμβανόμαστε το πεδίο. Βέβαια αυτό δεν συνεπάγεται και πλήρη στασιμότητα. Όπως είδαμε στην εργασία, και αυτό είναι κάτι που αναμένεται να συνεχιστεί, οι αλγόριθμοι και οι τεχνικές της Μηχανικής Μάθησης προσαρμόζονται στις νέες ανάγκες που επιτάσσει η εποχή. Ήδη έχει υπάρξει σημαντική πρόοδος ως προς τη μεταφορά της υλοποίησης σε κατανεμημένο περιβάλλον (όπως για παράδειγμα με το Apache Mahout [Lyub16]), ενώ αυτό που αναμένεται στο μέλλον είναι να προταθούν τρόποι για την αντιμετώπιση των προκλήσεων που εισάγει η ύπαρξη των ροών δεδομένων.

Βιβλιογραφία

[Alam13] M. S. Alam and S. T. Vuong, "Random Forest Classification for Detecting Android Malware", pp. 663–669, Aug 2013.

[Alam16] Furqan Alam, Rashid Mehmood, Iyad Katib and Aiiad Albeshri, "Analysis of Eight Data Mining Algorithms for Smarter Internet of Things (IoT)", *Procedia Computer Science*, vol. 98, pp. 437 – 442, 2016. The 7th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2016)/The 6th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2016)/Affiliated Workshops.

[Altm92] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression", *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[Amos16a] Brandon Amos, Bartosz Ludwiczuk and Mahadev Satyanarayanan, "OpenFace : A general-purpose face recognition library with mobile applications", 2016.

[Amos16b] Brandon Amos, Bartosz Ludwiczuk and Mahadev Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications", Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[Arth07] David Arthur and Sergei Vassilvitskii, "K-means++: the advantages of careful seeding", in *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[Baum66] Leonard E. Baum and Ted Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains", *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, 12 1966.

[Bono12] Flavio Bonomi, Rodolfo Milito, Jiang Zhu and Sateesh Addepalli, "Fog Computing and Its Role in the Internet of Things", in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, MCC '12, pp. 13–16, New York, NY, USA, 2012, ACM.

[Bove14] Gérôme Bovet, Antonio Ridi and Jean Hennebert, "Virtual Things for Machine Learning Applications", pp. 4–9, 2014.

[Cort95] Corinna Cortes and Vladimir Vapnik, "Support-vector networks", *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep 1995.

[Cram02] Jan Salomon Cramer, "The origins of logistic regression", 2002.

[DeCa09] Teófilo Emídio De Campos, Bodla Rakesh Babu, Manik Varma et al., "Character recognition in natural images.", *VISAPP (2)*, vol. 7, 2009.

[Demp77] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[Derg14] W. Derguech, E. Bruke and E. Curry, "An Autonomic Approach to Real-Time Predictive Analytics Using Open Data and Internet of Things", pp. 204–211, Dec 2014.

[Detr89] R Detrano, "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease", *American Journal of Cardiology*, vol. 64, pp. 304–310, 1989.

[Diro18] Abebe Abeshu Diro and Naveen Chilamkurti, "Distributed attack detection scheme using deep learning approach for Internet of Things", *Future Generation Computer Systems*, vol. 82, pp. 761 – 768, 2018.

[Dong11] Cao Dong, Xiuquan Qiao, Judith Gelernter, Li Xiaofeng and Meng Luoming, "Mining Data Correlation from Multi-faceted Sensor Data in the Internet of Things", vol. 8, 01 2011.

[Druc97] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola and Vladimir Vapnik, "Support vector regression machines", in *Advances in neural information processing systems*, pp. 155–161, 1997.

[Este96] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", pp. 226–231, AAAI Press, 1996.

[Fawc06] Tom Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006. ROC Analysis in Pattern Recognition.

[Feng17] S. Feng, P. Setoodeh and S. Haykin, “Smart Home: Cognitive Interactive People-Centric Internet of Things”, *IEEE Communications Magazine*, vol. 55, no. 2, pp. 34–39, February 2017.

[Freu96] Yoav Freund and Robert E. Schapire, “Experiments with a new boosting algorithm”, in *Thirteenth International Conference on Machine Learning*, pp. 148–156, San Francisco, 1996, Morgan Kaufmann.

[FRS01] Karl Pearson F.R.S., “LIII. On lines and planes of closest fit to systems of points in space”, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[Ganz15] F. Ganz, D. Puschmann, P. Barnaghi and F. Carrez, “A Practical Evaluation of Information Processing and Abstraction Techniques for the Internet of Things”, *IEEE Internet of Things Journal*, vol. 2, no. 4, pp. 340–354, Aug 2015.

[Gisl08] Drew Gislason, *Zigbee Wireless Networking*, Newnes, Newton, MA, USA, pap/onl edition, 2008.

[Gura17] Koppala Guravaiah, R. G. Thivyavignesh and R. Leela Velusamy, “Vehicle Monitoring Using Internet of Things”, pp. 24:1–24:7, 2017.

[Guve97] H. A. Guvenir, B. Acar, G. Demiroz and A. Cekin, “A supervised machine learning algorithm for arrhythmia analysis”, in *Computers in Cardiology 1997*, pp. 433–436, Sep. 1997.

[Han14] W. Han, Y. Gu, Y. Zhang and L. Zheng, “Data driven quantitative trust model for the Internet of Agricultural Things”, pp. 31–36, Oct 2014.

[Hass15] M. Hassanalieragh, A. Page, T. Soyata, G. Sharma, M. Aktas, G. Mateos, B. Kantarci and S. Andreescu, “Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-Based Processing: Opportunities and Challenges”, pp. 285–292, June 2015.

[Ho95] Tin Kam Ho, “Random Decision Forests”, in *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR ’95, pp. 278–, Washington, DC, USA, 1995, IEEE Computer Society.

[Hojs12] Søren Højsgaard, David Edwards and Steffen Lauritzen, *Gaussian Graphical Models*, pp. 77–116, Springer US, Boston, MA, 2012.

[Hoss16] M. Shamim Hossain and Ghulam Muhammad, "Cloud-assisted Industrial Internet of Things (IIoT) – Enabled framework for health monitoring", *Computer Networks*, vol. 101, pp. 192 – 202, 2016. Industrial Technologies and Applications for the Internet of Things.

[HOTE36] HAROLD HOTELLING, "RELATIONS BETWEEN TWO SETS OF VARIATES*", *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 12 1936.

[Hrom15] H. Hromic, D. Le Phuoc, M. Serrano, A. Antonić, I. P. Žarko, C. Hayes and S. Decker, "Real time analysis of sensor data for the Internet of Things by means of clustering and event processing", in *2015 IEEE International Conference on Communications (ICC)*, pp. 685–691, June 2015.

[Jakk10] V. Jakkula and D. Cook, "Outlier Detection in Smart Environment Structured Power Datasets", pp. 29–33, July 2010.

[Jin14] Jiong Jin, Jayavardhana Gubbi, Slaven Marusic and Marimuthu Palaniswami, "An Information Framework for Creating a Smart City Through Internet of Things", vol. 1, pp. 112–121, 04 2014.

[Kenn10] James Kennedy, "Particle swarm optimization", *Encyclopedia of machine learning*, pp. 760–766, 2010.

[Khan14] Muhammad Khan, Aunsia Khan, Muhammad Nasir Khan and Sajid Anwar, "A novel learning method to classify data streams in the internet of things", pp. 61–66, 11 2014.

[Kriz09] Alex Krizhevsky, Geoffrey Hinton et al., "Learning multiple layers of features from tiny images", Technical report, Citeseer, 2009.

[Kuma17] Ashish Kumar, Saurabh Goyal and Manik Varma, "Resource-efficient Machine Learning in 2 KB RAM for the Internet of Things", May 2017.

[Kuma18] Priyan Malarvizhi Kumar and Usha Devi Gandhi, "A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases", *Computers & Electrical Engineering*, vol. 65, pp. 222 – 235, 2018.

[LeCu10] Yann LeCun and Corinna Cortes, "MNIST handwritten digit database", 2010.

[LeCu15] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, "Deep learning", *nature*, vol. 521, no. 7553, p. 436, 2015.

[Lee16] J. Lee, M. Stanley, A. Spanias and C. Tepedelenlioglu, "Integrating machine learning in embedded sensor systems for Internet-of-Things applications", pp. 290–294, Dec 2016.

[Li11] Stan Z. Li and Anil K. Jain, *Handbook of Face Recognition*, Springer Publishing Company, Incorporated, 2nd edition, 2011.

[Li18] P. Li, Z. Chen, L. T. Yang, Q. Zhang and M. J. Deen, "Deep Convolutional Computation Model for Feature Learning on Big Data in Internet of Things", *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 790–798, Feb 2018.

[Lyub16] Dmitriy Lyubimov and Andrew Palumbo, *Apache Mahout: Beyond MapReduce*, CreateSpace Independent Publishing Platform, USA, 1st edition, 2016.

[Ma13] Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen and Jianfeng Liu, "Mining smart card data for transit riders' travel patterns", *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1 – 12, 2013.

[Ma15] Xiaolei Ma, Haiyang Yu, Yunpeng Wang and Yinhai Wang, "Large-scale transportation network congestion evolution prediction using deep learning theory", *PloS one*, vol. 10, no. 3, p. e0119044, 2015.

[MacQ67] J. MacQueen, "Some methods for classification and analysis of multivariate observations", in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 281–297, Berkeley, Calif., 1967, University of California Press.

[Mano18] Gunasekaran Manogaran, R. Varatharajan, Daphne Lopez, Priyan Malarvizhi Kumar, Revathi Sundarasekar and Chandu Thota, "A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system", *Future Generation Computer Systems*, vol. 82, pp. 375 – 387, 2018.

[Maro61] M. E. Maron, "Automatic Indexing: An Experimental Inquiry", *J. ACM*, vol. 8, no. 3, pp. 404–417, July 1961.

[McCa00] Andrew McCallum, Kamal Nigam and Lyle H. Ungar, "Efficient Clustering of High-dimensional Data Sets with Application to Reference Matching", in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pp. 169–178, New York, NY, USA, 2000, ACM.

[McRo10] Michael McRoberts, *Beginning Arduino*, Apress, Berkely, CA, USA, 1st edition, 2010.

[Mone13] Dorothy N. Monekosso and Paolo Remagnino, "Data reconciliation in a smart home sensor network", *Expert Systems with Applications*, vol. 40, no. 8, pp. 3248 – 3255, 2013.

[Ni14] P. Ni, C. Zhang and Y. Ji, "A hybrid method for short-term sensor data forecasting in Internet of Things", pp. 369–373, Aug 2014.

[Nick08] John Nickolls, Ian Buck, Michael Garland and Kevin Skadron, "Scalable Parallel Programming with CUDA", *Queue*, vol. 6, no. 2, pp. 40–53, March 2008.

[Outc17] Aissam Outchakoucht, Hamza ES-SAMAALI and Jean Philippe, "Dynamic Access Control Policy based on Blockchain and Machine Learning for the Internet of Things", vol. 8, 01 2017.

[Pand17] P. S. Pandey, "Machine Learning and IoT for prediction and detection of stress", pp. 1–5, July 2017.

[Pasl15] Cristian Pasluosta, Heiko Gaßner, Juergen Winkler, Jochen Klucken and Bjoern Eskofier, "An Emerging Era in the Management of Parkinson's Disease: Wearable Technologies and the Internet of Things", vol. 19, 07 2015.

[Pawl98] Zdzislaw Pawlak, "Rough set theory and its applications to data analysis", *Cybernetics & Systems*, vol. 29, no. 7, pp. 661–688, 1998.

[Pear85] Judea Pearl, "Bayesian netwcrks: A model cf self-activated memory for evidential reasoning", 1985.

[Pele00] David Peleg, "Distributed computing", *SIAM Monographs on discrete mathematics and applications*, vol. 5, pp. 1–1, 2000.

[Rahm15] Amir M. Rahmani, Nanda Kumar Thanigaivelan, Tuan Nguyen gia, Jose Granados, Behailu Shiferaw Negash, Pasi Liljeberg and Hannu Tenhunen, "Smart e-Health Gateway: Bringing Intelligence to Internet-of-Things Based Ubiquitous Healthcare Systems", 07 2015.

[Rath16] M. Mazhar Rathore, Awais Ahmad, Anand Paul and Seungmin Rho, "Urban planning and building smart cities based on the Internet of Things using Big Data analytics", *Computer Networks*, vol. 101, pp. 63 – 80, 2016. Industrial Technologies and Applications for the Internet of Things.

[Reyn15] Douglas Reynolds, "Gaussian mixture models", *Encyclopedia of biometrics*, pp. 827–832, 2015.

[Ruta18] Michele Ruta, Floriano Scioscia, Giuseppe Loseto, Agnese Pinto and Eugenio Di Sciascio, "Machine learning in the Internet of things: A semantic-enhanced approach", pp. 1–22, 08 2018.

[Sanc14] Luis Sanchez, Luis Muñoz, Jose Antonio Galache, Pablo Sotres, Juan R. Santana, Veronica Gutierrez, Rajiv Ramdhany, Alex Gluhak, Srdjan Krco, Evangelos Theodoridis and Dennis Pfisterer, "SmartSantander: IoT experimentation over a smart city testbed", *Computer Networks*, vol. 61, pp. 217 – 238, mar 2014. Special issue on Future Internet Testbeds – Part I.

[Scho99] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor and John Platt, "Support Vector Method for Novelty Detection", in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pp. 582–588, Cambridge, MA, USA, 1999, MIT Press.

[Shuk15] M. Shukla, Y. P. Kosta and P. Chauhan, "Analysis and evaluation of outlier detection algorithms in data streams", pp. 1–8, Sept 2015.

[Somo13] A. Somov, C. Dupont and R. Giaffreda, "Supporting smart-city mobility with cognitive Internet of Things", pp. 1–10, July 2013.

[Souz15] Alberto M.C. Souza and Joseé R.A. Amazonas, "An Outlier Detect Algorithm using Big Data Processing and Internet of Things Architecture", *Procedia Computer Science*, vol. 52, pp. 1010 – 1015, 2015. The 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015).

[Swan15] Melanie Swan, *Blockchain: Blueprint for a New Economy*, O'Reilly Media, Inc., 1st edition, 2015.

[Tava09] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", in *Proceedings of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications*, CISDA'09, pp. 53–58, Piscataway, NJ, USA, 2009, IEEE Press.

[Uzel15] Ana Uzelac, Nenad Gligoric and Srdjan Krco, "A comprehensive study of parameters in physical environment that impact students' focus during lecture

using Internet of Things”, *Computers in Human Behavior*, vol. 53, pp. 427 – 434, 2015.

[Vavi13] Vinod Kumar Vavilapalli, Arun C. Murthy, Chris Douglas, Sharad Agarwal, Mahadev Konar, Robert Evans, Thomas Graves, Jason Lowe, Hitesh Shah, Siddharth Seth, Bikas Saha, Carlo Curino, Owen O’Malley, Sanjay Radia, Benjamin Reed and Eric Baldeschwieler, “Apache Hadoop YARN: Yet Another Resource Negotiator”, in *Proceedings of the 4th Annual Symposium on Cloud Computing*, SOCC ’13, pp. 5:1–5:16, New York, NY, USA, 2013, ACM.

[WU09] ZHAOHUA WU and NORDEN E. HUANG, “ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD”, *Advances in Adaptive Data Analysis*, vol. 01, no. 01, pp. 1–41, 2009.

[Yang] Allen Y. Yang, Philip Kuryloski and Ruzena Bajcsy, “WARD: A Wearable Action Recognition Database”.

[Yin13] Y. Yin and D. Jiang, “Research and Application on Intelligent Parking Solution Based on Internet of Things”, vol. 2, pp. 101–105, Aug 2013.

[Yogi13] Yogita and D. Toshniwal, “Clustering techniques for streaming data-a survey”, pp. 951–956, Feb 2013.

[Zaha16] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker and Ion Stoica, “Apache Spark: A Unified Engine for Big Data Processing”, *Commun. ACM*, vol. 59, no. 11, pp. 56–65, October 2016.

[Zane12] Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista and Michele Zorzi, “Internet of Things for Smart Cities”, vol. 1, 01 2012.

Παράρτημα Α

Ευρετήριο Ακρωνυμίων και Συντμήσεων

EEMD: Ensemble Empirical Mode Decomposition

IoT: Internet of Things (Διαδίκτυο των Πραγμάτων)

kNN: k-Nearest Neighbors (k-Πλησιέστεροι Γείτονες)

LR: Logistic Regression (Λογιστική Παλινδρόμηση)

ML: Machine Learning (Μηχανική Μάθηση)

MSE: Mean Squared Error (Μέσο Τετραγωνικό Σφάλμα)

OCSVM: One-class Support Vector Machines (Μηχανές Διανυσμάτων Υποστήριξης μίας Κλάσης)

PSO: Particle Swarm Optimization

ROC: Receiver Operating Characteristic (Λειτουργικό Χαρακτηριστικό Δέκτη)

SVM: Support Vector Machines (Μηχανές Διανυσμάτων Υποστήριξης)

SVR: Support Vector Regression (Παλινδρόμηση Διανυσμάτων Υποστήριξης)