



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΓΕΝΙΚΟ ΤΜΗΜΑ, ΛΑΡΙΣΑ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΕΝΕΡΓΕΙΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΣΥΣΤΗΜΑΤΑ ΑΥΤΟΜΑΤΙΣΜΩΝ»

Εισαγωγική μελέτη των εφαρμογών της
Ερμηνεύσιμης Τεχνητής Νοημοσύνης στη
Βιομηχανία

Introduction to Explainable Artificial Intelligence Applications in Industry

4.0

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παπαγιάννης Κωνσταντίνος

ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ: Δρ. ΠΑΠΑΓΕΩΡΓΙΟΥ ΕΛΠΙΝΙΚΗ

ΛΑΡΙΣΑ, ΟΚΤΩΒΡΙΟΣ 2023

Υπεύθυνη Δήλωση

«Δηλώνω υπεύθυνα ότι η συγκεκριμένη μεταπτυχιακή διπλωματική εργασία για τη λήψη του μεταπτυχιακού τίτλου σπουδών του ΠΜΣ Πλήρους Φοίτησης του Πανεπιστημίου Θεσσαλίας «Ενεργειακές Τεχνολογίες και Συστήματα Αυτοματισμών» έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει εγκριθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό. Η εργασία αυτή έχοντας εκπονηθεί από εμένα, αντιπροσωπεύει τις προσωπικές μου απόψεις επί του θέματος και το κείμενο είναι γραμμένο με τα δικά μου λόγια και δεν αποτελεί προϊόν λογοκλοπής από τρίτες πηγές. Οι πηγές στις οποίες ανέτρεξα για την εκπόνηση της συγκεκριμένης διπλωματικής αναφέρονται στο σύνολό τους, δίνοντας πλήρεις αναφορές στους συγγραφείς, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο».

Ο Δηλών

Περίληψη

Η διπλωματική αυτή υλοποιήθηκε από τον σπουδαστή του Μεταπτυχιακού Προγράμματος Σπουδών «Ενεργειακές Τεχνολογίες και Συστήματα Αυτοματισμών» Παπαγιάννη Κωνσταντίνο. Το θέμα που παρουσιάζεται είναι «Εισαγωγική μελέτη των εφαρμογών της Ερμηνεύσιμης Τεχνητής Νοημοσύνης στη Βιομηχανία». Επιβλέπουσα της παρούσας εργασίας είναι η καθηγήτρια του Π.Μ.Σ. Ελπινίκη Παπαγεωργίου.

Η τεχνητή νοημοσύνη στην έξυπνη βιομηχανία είναι αδιαμφισβήτητα ένας πολλά υποσχόμενος και εξελίξιμος κλάδος. Η αύξηση της παραγωγής, η διασφάλιση της ασφάλειας των εργαζομένων και της ποιότητας των προϊόντων, καθώς επίσης και η βελτίωση των συνθηκών εργασίας, είναι μερικές από τις συνεισφορές της τεχνητής νοημοσύνης στον συγκεκριμένο κλάδο. Παρόλα αυτά τα περισσότερα μοντέλα τεχνητής νοημοσύνης που αναπτύσσονται λειτουργούν σαν «μαύρα κουτιά», με την έννοια ότι ο τρόπος λειτουργίας τους δεν είναι διαφανής ως προς τον χρήστη της. Για τον λόγο αυτό τα επεξηγηματικά μοντέλα τεχνητής νοημοσύνης, που θα λειτουργούν σαν «λευκά κουτιά» έχουν ξεκινήσει να είναι το αντικείμενο αιχμής στον κλάδο.

Στόχος της παρούσας διπλωματικής είναι να εξεταστούν μεθοδολογίες, προτερήματα και μειονεκτήματα των επεξηγηματικών μοντέλων, και η εφαρμογή τους στην έξυπνη βιομηχανία. Βάσει της διεθνούς βιβλιογραφίας, αναλύεται η απαραίτητη θεωρία και η εφαρμογή των επεξηγηματικών μοντέλων στην έξυπνη βιομηχανία. Η δομή της συγκεκριμένης εργασίας έχει ως στόχο να εισάγει την απαραίτητη θεωρία στον αναγνώστη μέσα από τα πρώτα εισαγωγικά κεφάλαια της Βιομηχανίας 4.0 (Κεφάλαιο 1), της Μηχανικής και της Βαθιάς Μάθησης (Κεφάλαιο 2), και της Επεξηγηματική Τεχνητής Νοημοσύνης (Κεφάλαιο 3) ώστε να γίνει πιο βαθιά η μετάβαση στο κεφάλαιο ανάλυσης των μεθοδολογιών που βρέθηκαν στην βιβλιογραφική ανασκόπηση (Κεφάλαιο 4). Ιδιαίτερη έμφαση δίνεται σε εφαρμογές μηχανικής και βαθιάς μάθησης. Τα ευρήματα της εργασίας δηλώνουν ότι η επεξηγηματική τεχνητή νοημοσύνη έχει ήδη αρχίσει να εφαρμόζεται σε αρκετούς τομείς της έξυπνης βιομηχανίας. Επίσης συμπεραίνεται ότι ο κλάδος της επεξηγηματικής τεχνητής έχει πολλά ακόμα να προσφέρει στην έξυπνη βιομηχανία και υπάρχει αρκετό κενό που μπορεί να καλυφθεί μέσω της έρευνας.

Λέξεις Κλειδί: Τεχνητή Νοημοσύνη, Επεξηγηματική Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Βαθιά Μάθηση

Abstract

This dissertation was created by the postgraduate student Papagiannis Konstantinos of the MSc. Program in “Introduction to Explainable Artificial Intelligence Applications in Industry 4.0” of the University of Thessaly’s General Department, on the topic of “Explainable Artificial Intelligence models for smart manufacturing”, under the supervision of the Professor Dr. Elpiniki Papageorgiou.

Artificial Intelligence (AI) in smart manufacturing is a growing and promising research field with lot of practical potentials. Through AI, smart manufacturing can benefit in areas such as the production optimization, the work-safety, the automatic quality inspection etc. However, most of the AI models that are currently implemented, can be considered as “black box” models. This practically means, that the user is not able to understand the working principle and the decision-making process of these models. Hence, efforts of the research community are being focused on Explainable Artificial Intelligence (XAI) models to alleviate and convert the “black box” principle into “white box”.

The aim of this work is to discover XAI methodologies that are developed for smart manufacturing, and to further address their pros and cons. Based on the literature findings, the basic theory, and the appliances of XAI models are presented. The structure of this dissertation aims to introduce the basic concepts to the reader with the first theoretical sections though the sections 1-3 that present the aspects of XAI, in order to help him understand the methodologies that are presented in section 4. This work mostly focuses on machine learning and deep learning methods that were found in literature. The findings of this work suggest that the XAI has been already implemented in a variety of smart manufacturing applications. Nevertheless, it is concluded that XAI has still a lot of potentials and offers a great opportunity for further research in smart manufacturing.

Keywords: Artificial Intelligence, Explainable Artificial Intelligence, Machine Learning, Deep Learning

Ευχαριστίες

Η ολοκλήρωση των σπουδών, μου με την υλοποίηση αυτής της διπλωματικής, είναι ένα χαρμόσυνο αλλά ταυτόχρονα και νοσταλγικό γεγονός. Μέσα από το Π.Μ.Σ. διεύρυνα τις γνώσεις μου και ανέπτυξα φιλικές και επαγγελματικές σχέσεις με τους συμφοιτητές μου. Για το λόγο αυτό θέλω να ευχαριστήσω τους καθηγητές και τους συμφοιτητές μου για την άψογη συνεργασία. Ιδιαίτερα θέλω να ευχαριστήσω την επιβλέπων καθηγήτρια Παπαγεωργίου Ελπινίκη και τον συνάδελφο μου Θεόδωρο Τζιόλα, συμφοιτητή στο Π.Μ.Σ., καθώς συνέβαλε και αυτός με την πολύτιμη βοήθεια και καθοδήγηση του στην ολοκλήρωση αυτής της διπλωματικής.

Περιεχόμενα

Περίληψη.....	4
Abstract	5
Ευχαριστίες.....	6
Ευρετήριο εικόνων και πινάκων	8
Εισαγωγή	10
1. Εισαγωγή στη Βιομηχανία 4.0	10
1.1. Έξυπνα εργοστάσια και κατασκευές	12
1.2. Αλληλεπίδραση ανθρώπου-μηχανής	13
1.3. Προληπτική συντήρηση	13
1.4. Έξυπνα προϊόντα	14
1.5. Βιομηχανική Ρομποτική	14
1.6. Ασφάλεια στον κυβερνοχώρο και ιδιωτικότητα	14
1.7. Έξυπνες Μεταφορές	14
2. Μηχανική και Βαθιά Μάθηση	15
3. Επεξηγηματική Τεχνητή Νοημοσύνη	17
3.1. Κατηγορίες Μεθόδων Επεξήγησης	20
3.2. Διαφανή μοντέλα και σύνοψη των Post-hoc τεχνικών ETN.	23
3.3. Μετρικά επίδοσης - αξιολόγησης	28
4. Μέθοδοι Επεξηγηματικών Τεχνικών που χρησιμοποιήθηκαν στη βιομηχανία	31
5. Επίλογος και συμπεράσματα	37
Πηγές.....	38

Ευρετήριο εικόνων και πινάκων

Εικόνα 1 Σχηματική απεικόνιση των τεσσάρων βιομηχανικών επαναστάσεων [3].....	11
Εικόνα 2 Βασικές τεχνολογίες γενικής εφαρμογής στη Βιομηχανία 4.0 [3].	12
Εικόνα 3 Η αρχική εικόνα (x) που δείχνει ένα ζώο πάντα, δέχεται σκόπιμα θόρυβο στον αλγόριθμο αντιμαχόμενων δικτύων για να ξεγελάσει το δίκτυο που έχει ως στόχο να ταξινομήσει σωστά τις εικόνες. Με την εισαγωγή θορύβου το δίκτυο ξεγελιέται και δίνει μεγάλη σιγουριά (99.3%) ότι είναι εικόνα της κλάσης γίββων, αν και το αποτέλεσμα στο ανθρώπινο μάτι δεν έχει αλλάξει. Πηγή [15]. ..	19
Εικόνα 4 Οι 4 κατηγορίες ταξινόμησης μεθόδων ETN όπως ορίζονται στην [3].....	20
Εικόνα 5 Παράδειγμα εφαρμογής του Deep SHAP στο γνωστό πακέτο δεδομένων αριθμών γραφικού χαρακτήρα MNIST. Με κόκκινο χρώμα οι περιοχές pixels που συνεισφέρουν σε μια κλάση ενώ με μπλε αυτές οι περιοχές που δεν συνεισφέρουν σε μια κλάση.	22
Εικόνα 6 Ταξινόμηση σύμφωνα με την [19] για τις τεχνικές ETN που βρέθηκαν στην βιβλιογραφία. Χρησιμοποιούν χρωματισμό στις παραπομπές που εμφανίζουν όπως μπλε, πράσινο και κόκκινο που αντιστοιχούν σε ETN τεχνικές που χρησιμοποιούν εικόνες, κείμενο ή δεδομένα σε μορφή πίνακα. Για την ταξινόμηση αυτή έψαξαν αναλυτικά στην βιβλιογραφία ώστε να συμπεριλάβουν μεθόδους που μπορούν να χρησιμοποιηθούν σε όλα τα μοντέλα.	26
Εικόνα 7 Παράδειγμα της οπτικοποίησης των αποτελεσμάτων των μεθόδων που προτάθηκαν στην [27]. Αριστερά (a) φαίνεται ξεκάθαρα η περιοχή που το μοντέλο αναγνώρισε με ελάττωμα, ενώ δεξιά (b) κάποιες περιοχές που λανθασμένα ενεργοποιήθηκαν.	33
Εικόνα 8 Η εφαρμογή της μεθοδολογίας 3D Grad-CAM που φανερώνει της περιοχές με δυσκολία κατασκευής [28].....	33
Εικόνα 9 Τα αποτελέσματα της ETN μεθόδου Grad-CAM που χρησιμοποίησαν στην [30].	34
Εικόνα 10 Σύγκριση της τεχνικής LIME με την VAE-LIME για διαφορές μεταβλητές της εργασίας [31].	35
Εικόνα 11 Διάγνωση με την χρήση ETN για την ίδια καταγραφή πρεσαρίσματος με 4 τύπους δυσλειτουργίας, σκιαγραφίζοντας με κόκκινο χρώμα τα σημεία όπου δημιουργούν τις δυσλειτουργίες.	36

Πίνακας 1 Ταξινόμηση τεχνικών ETN βάση της [1] με ποικίλα κριτήρια.....	27
---	----

Εισαγωγή

Η τεχνητή νοημοσύνη (TN) είναι ένας κλάδος της επιστήμης των υπολογιστών που ασχολείται με τη δημιουργία και την εφαρμογή μηχανών και λογισμικών που μπορούν να εκτελέσουν εργασίες που απαιτούν νοημοσύνη, όπως η αναγνώριση προτύπων, η επίλυση προβλημάτων, η μάθηση και η δημιουργικότητα. Η TN είναι ένας πολύ ευρύς και δυναμικός τομέας, που συνεχώς εξελίσσεται και βελτιώνεται με τη χρήση νέων δεδομένων, αλγορίθμων και υπολογιστικής ισχύος [1]. Η αυξανόμενη ψηφιοποίηση κάθε πτυχής της ζωής παρέχει τεράστιες ποσότητες δεδομένων, επιτρέποντας την εφαρμογή μοντέλων TN.

Οι κλάδοι της βιομηχανίας δεν αποτελούν εξαίρεση στην τάση αυτή. Τα TN παίζουν σημαντικό ρόλο σε πολλές πτυχές της παραγωγικής διαδικασίας καθώς βελτιώνουν την ποιοτικό έλεγχο στις γραμμές παραγωγής, την εύκολη αναδιαμόρφωση και προσαρμογή του αυτοματοποιημένου χειρισμού εξαρτημάτων, την διάγνωση σφαλμάτων και πρόβλεψη συμβάντων, την πιο ευέλικτη διαχείριση παραγωγής, τον ευέλικτο σχεδιασμό παραγωγής και τη δυνατότητα ασφαλούς συνεργασίας μεταξύ ανθρώπων και ρομπότ. Ειδικά η συνεργασία μεταξύ ανθρώπων και ρομπότ, είναι πρώτιστης σημασίας για τη μετάβαση στην έξυπνη βιομηχανία.

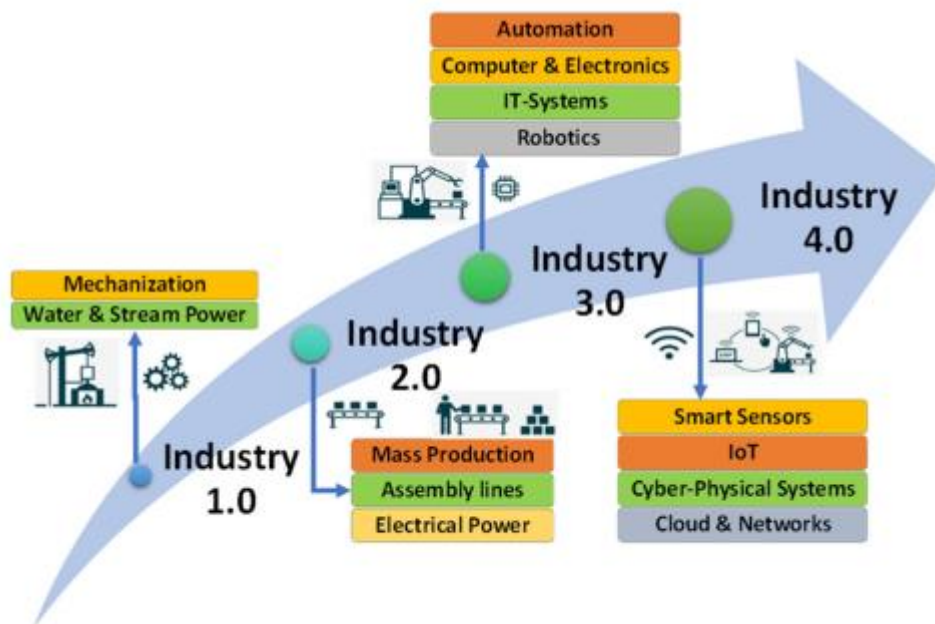
Τα μοντέλα τεχνητής νοημοσύνης παρέχουν τα μέσα για την αυτοματοποίηση πολλών εργασιών και την επίτευξη επιπέδων απόδοσης χωρίς προηγούμενο. Ωστόσο, στις περισσότερες περιπτώσεις, τα μοντέλα αυτά είναι αδιαφανή για τον χρήστη: δηλαδή λειτουργούν ως «μαύρα κουτιά». Οι προβλέψεις τους, αν και είναι ακριβείς, δεν φανερώνουν την συλλογιστική διαδικασία στους χρήστες. Δεδομένης της επίδρασης αυτών των προβλέψεων στις διαδικασίες λήψης αποφάσεων, είναι ζωτικής σημασίας η ανάπτυξη μηχανισμών και τεχνικών για την παροχή πληροφοριών στους χρήστες σχετικά με μια τέτοια διαδικασία συλλογισμού μοντέλων TN. Η ανάπτυξη αυτών των τεχνικών και μηχανισμών και ο τρόπος παρουσίασης αυτών των ιδεών έχει γεννήσει ένα δικό του ερευνητικό πεδίο, γνωστό ως Επεξηγηματική Τεχνητή Νοημοσύνη (ETN). Ενώ το πεδίο της ETN μπορεί να αναχθεί στη δεκαετία του 1970 [2], τα τελευταία χρόνια έχει γνωρίσει μια αυξημένη ζήτηση. Σε αυτό έχει συμβάλει ως επί τω πλείστων η ραγδαία στροφή προς την μηχανική και τη βαθιά μάθηση (Machine Learning, Deep Learning) που είναι ένα από τα βασικά στοιχεία της τέταρτης βιομηχανικής επανάστασης που διανύουμε στην εποχή μας, ή όπως είναι γνωστή με τον όρο Βιομηχανία 4.0 [3] και η έννοια της περιγράφεται σε επόμενο κεφάλαιο.

Πιο αναλυτικά, στα επόμενα κεφάλαια παρουσιάζονται μεθοδολογίες και τεχνικές ETN που εφαρμόστηκαν σε βιομηχανικές λειτουργίες. Έμφαση δίνεται σε προηγούμενες βιβλιογραφικές ανασκοπήσεις, και συγκεκριμένα στις [1], [3], [4] που βρέθηκαν στην βάση δεδομένων Google Scholar χρησιμοποιώντας τις λέξεις κλειδί της διεθνούς βιβλιογραφίας: Explainable Artificial Intelligence, Explainable Deep Learning Models, Explainable Machine Learning. Σαν κριτήρια αναζήτησης, επιλέχθηκε η ημερομηνία δημοσίευσης. Συγκεκριμένα επιλέχθηκαν ανασκοπήσεις από το 2020 και μετά (τελευταίας τριετίας από την ημερομηνία συγγραφής αυτής της εργασίας). Η δομή που ακολουθείται περιλαμβάνει μια θεωρητική εισαγωγή στις έννοιες: Βιομηχανία 4.0, Μηχανική Μάθηση (ML), Βαθιά Μάθηση (DL), Επεξηγηματική Τεχνητή Νοημοσύνη (ETN). Εν συνεχεία, περιγράφονται οι πιο δημοφιλείς τεχνικές και μέθοδοι ETN και η εφαρμογή τους σε βιομηχανικά προβλήματα. Τέλος εξάγονται χρήσιμα συμπεράσματα.

1. Εισαγωγή στη Βιομηχανία 4.0

Ο πρώτος βιομηχανικός μετασχηματισμός άρχισε με τη χρήση ατμού και την ενέργεια του νερού, ενώ ο δεύτερος βιομηχανικός μετασχηματισμός ξεκίνησε μαζική παραγωγή μέσω ηλεκτρισμού. Η τρίτη επανάσταση έθεσε τη συμμετοχή των υπολογιστών και των ηλεκτρονικών ειδών στη βιομηχανία; η

επικείμενη βιομηχανική επανάσταση, που ονομάζεται Βιομηχανία 4.0, ενσωματώνει την αυτοματοποίηση και τις ψηφιακές τεχνολογίες δίδυμων (digital twin) όπως απεικονίζεται στην Εικόνα 1. Ως εκ τούτου, ο κλάδος 4.0 έχει ερμηνευθεί ως «η νέα κατεύθυνση της αυτοματοποίησης και της μεταφοράς ψηφιακών δεδομένων στην κατασκευή και παρόμοιων τεχνολογιών, συμπεριλαμβανομένου του Διαδικτύου των Πραγμάτων (Internet of Things ή εν συντομία IoT), των κυβερνο-φυσικών συστημάτων, του υπολογιστικού νέφους, της ενοποίησης συστημάτων, της ανάλυσης μεγάλων δεδομένων που παρέχει υπηρεσίες για τη δημιουργία των έξυπνων βιομηχανιών και εργοστασίων». Η Βιομηχανία 4.0 περιλαμβάνει το ευφύες δίκτυο συσκευών, μηχανημάτων και συστημάτων για βιομηχανίες με τη βοήθεια της επικοινωνίας και της τεχνολογίας. Η πρόοδος της ΤΝ και η χρήση μεθόδων μηχανικής μάθησης (ML) και βαθιάς μάθησης (DL) σε βιομηχανικά προβλήματα αποτελούν μέρος της Βιομηχανίας 4.0. Οι εφαρμογές ΤΝ έχουν σημειώσει πρόοδο στην επίλυση της αυτόματης αναγνώρισης των μοτίβων στα δεδομένα. Τα συστήματα που βασίζονται στην ΤΝ μπορούν να βοηθήσουν τους ειδικούς καθώς πραγματοποιούν αξιολογήσεις στο πλαίσιο των δραστηριοτήτων τους. Η στοχευμένη εφαρμογή των δεδομένων και της λήψης αποφάσεων μπορεί να επηρεάσει σημαντικά την αύξηση της παραγωγικότητας στους τομείς της βιομηχανίας, επιτρέποντας την επιτυχή λειτουργία και την ενσωμάτωση γνώσεων στις επιχειρηματικές διαδικασίες. Η ενσωμάτωση αυτή χρειάζεται μια προηγμένη διαδικασία διαχείρισης, η οποία αναπτύσσει την εμπιστοσύνη στις ενέργειες, τους μηχανισμούς συμπερασμάτων και τα αποτελέσματα των εκτεταμένων συστημάτων που βασίζονται στην τεχνητή νοημοσύνη.



Εικόνα 1 Σχηματική απεικόνιση των τεσσάρων βιομηχανικών επαναστάσεων [3].

Η Βιομηχανία 4.0 αντιπροσωπεύει μια δημοφιλή έννοια στον κατασκευαστικό και βιομηχανικό τομέα. Οι αρχικοί στόχοι συνήθως περιλαμβάνουν την αυτοματοποίηση, την ανάπτυξη της διαδικασίας και τη βελτίωση της παραγωγικότητας, οι πιο προηγμένοι στόχοι επικεντρώνονται στον εκσυγχρονισμό, τη λειτουργικότητα και την εξέλιξη διαφορετικών ευφύων συστημάτων ή μοντέλων με πληροφορίες και συνεργασίες ως απαραίτητα στοιχεία. Η κύρια ιδέα της Βιομηχανίας 4.0 είναι να παράγει εξαιρετική λειτουργική απόδοση και παραγωγικότητα και ταυτόχρονα ένα ισχυρό επίπεδο αυτοματοποίησης. Επιπλέον, η βιομηχανία 4.0 σκοπεύει να παρέχει ευφυή, σε πραγματικό χρόνο, έξυπνα, λειτουργικά και ανεξάρτητα βιομηχανικά περιβάλλοντα και πλατφόρμες. Για την επίτευξη του στόχου αυτού, βελτιωμένες και καινοτόμες διαδικασίες όπως [5]: η εξόρυξη πληθώρας δεδομένων (Big Data),

τεχνολογίες IoT, υπολογιστικού νέφους (cloud computing), κυβερνο-φυσικά συστήματα (cyber-physical systems), ενοποιημένα συστήματα (system integration), αυτόνομα και ρομποτικά συστήματα (autonomous robotics), προσομοίωση (simulation), και η παρασκευή τρισδιάστατων προϊόντων (additive manufacturing), ενσωματώνονται στο βιομηχανικό λογισμικό, όπως παρουσιάζεται στην Εικόνα 2.



Εικόνα 2 Βασικές τεχνολογίες γενικής εφαρμογής στη Βιομηχανία 4.0 [3].

Η τεχνητή νοημοσύνη στη Βιομηχανία 4.0 έχει το ρόλο να ενσωματώνει διάφορες τεχνολογίες που επιτρέπουν στο λογισμικό, τα συστήματα, τις μηχανές και τις συσκευές να ανιχνεύουν, να αντιλαμβάνονται, να αναπτύσσουν, να κατανοούν και να μαθαίνουν ή να διευρύνουν τις ανθρώπινες δραστηριότητες. Με την τεχνητή νοημοσύνη, τα βιομηχανικά συστήματα παραγωγής μπορούν να αποκτήσουν πιο εξαιρετικές επιδόσεις από τους ανθρώπους. Επιπλέον, η τεχνητή νοημοσύνη μπορεί να επιτρέψει στα ρομπότ να ολοκληρώνουν εργασίες που ένας άνθρωπος δεν θα εκτελούσε, όπως για παράδειγμα, να διαχειρίζεται κρίσιμες ή επικίνδυνες πρώτες ύλες ή μικροσκοπικά στοιχεία. Φυσικά είναι αναγκαίο για να τονιστεί ότι αυτήν τη στιγμή πολλά βιομηχανικά ρομπότ δεν είναι τόσο έξυπνα όσο οι άνθρωποι. Ωστόσο, μπορούν πράγματι να προσφέρουν πολλές επιδέξιες εργασίες. Η συνεχής πρόοδος κατέδειξε ότι αυτή η τεχνολογία εξελίσσεται με τη Βιομηχανία 4.0.

1.1. Έξυπνα εργοστάσια και κατασκευές

Η βιομηχανία 4.0 μεταμορφώνει τα εργοστάσια μέσα από καινοτόμες μεθόδους που υπόσχονται να τα μετατρέψουν σε ευέλικτα, πλήρως συνδεδεμένα, ανταποκρινόμενα και εξειδικευμένα στις ανάγκες της παραγωγής και του ανταγωνισμού. Μία πληθώρα νέων εφαρμογών μπορούν να χρησιμοποιηθούν και να ερευνηθούν περαιτέρω όπως το IoT και η TN. Τα έξυπνα εργοστάσια καθιέρωσαν με επιτυχία έναν ψηφιακό μετασχηματισμό και μπορούν να δημιουργήσουν και να προωθήσουν τα σύνολα δεδομένων τους για να εφαρμόσουν TN και ETN για να βελτιώσουν τη διαχείριση ποιότητας, την κανονικότητα και τη συντήρηση του εξοπλισμού αναπτύσσοντας προγνωστικές μελέτες της λειτουργικότητάς του και δίνουν πλήρη λογική στα όρια του εργοστασίου. Ενώ τα πλεονεκτήματα της TN είναι δημοφιλή όταν

αυτή εφαρμόζεται σε διαδικασίες παραγωγής, είναι απαραίτητο να προσδιοριστεί ότι τα εργοστάσια θα πρέπει να έχουν ένα σχέδιο ανάπτυξης τόσο TN και ETN όσο και ένα σύστημα για το είδος της βιομηχανικής και αυτοματοποιημένης πλατφόρμας που θα εφαρμοστούν τα νέα εργαλεία. Η εισαγωγή συσκευών IoT [5], μεγάλων δεδομένων και υπολογιστικού νέφους [6], καθώς και τα κυβερνο-φυσικά συστήματα [7] στα εργοστάσια είναι η πιο αποτελεσματική μέθοδος για τη μετατροπή συμβατικών εργοστασίων σε έξυπνα εργοστάσια που θα ανήκουν στη βιομηχανία 4.0.

Η βιομηχανία 4.0 προωθεί επίσης την κατασκευή εργοστασίων πιο έξυπνων, προσαρμόσιμων και παραγωγικών, εφαρμόζοντας και εξοπλίζοντάς τα με έξυπνους αισθητήρες, συσκευές και ανεξάρτητα, αυτόνομα συστήματα. Κατά συνέπεια, οι μηχανές, τα συστήματα και ο εξοπλισμός παράγουν υψηλά πρότυπα αυτο-βελτιστοποίησης και εκβιομηχάνισης. Επιπλέον, η διαδικασία παραγωγής περιλαμβάνει την ικανότητα εκπλήρωσης περαιτέρω σύνθετων και αποδοτικών προδιαγραφών και συνθηκών των προϊόντων. Ως εκ τούτου, τα έξυπνα εργοστάσια και η έξυπνη κατασκευή είναι οι κύριοι στόχοι της βιομηχανίας 4.0. Όταν υλοποιούνται με την κατάλληλη προσέγγιση, η τεχνητή νοημοσύνη και η επεξηγηματική τεχνητή νοημοσύνη έχουν πολλά οφέλη για την κατασκευαστική βιομηχανία, συμπεριλαμβανομένης της μείωσης των σφαλμάτων (οι ευφυείς αλγόριθμοι μπορούν να εκτελέσουν εργασίες και να μειώσουν την ευαισθησία σε σφάλματα σε διαδικασίες που εκτελούνται από ανθρώπους), τη μείωση του κόστους (διάφορες e-commerce shops) οι τράπεζες χρησιμοποιούν ρομπότ για την βοήθεια των καταναλωτών. Με αυτόν τον τρόπο, οι οργανισμοί μπορούν να μειώσουν τα έξοδα των εργαζομένων ή να τα αναθέσουν σε άλλες θέσεις εργασίας σε πιο επιτακτικούς και σύνθετους τομείς που μπορούν να ενισχύσουν το κέρδος της και την ποιότητα εργασίας.

1.2. Αλληλεπίδραση ανθρώπου-μηχανής

Σε αυτήν την ενότητα αναλύονται οι τρόποι με τους οποίους οι άνθρωποι αλληλοεπιδρούν με μηχανήματα και συστήματα. Στην Βιομηχανία 4.0., η βελτιωμένη και αναπτυσσόμενη αλληλεπίδραση ανθρώπου-μηχανής προσπαθεί να απαντήσει στο ερώτημα: πώς μπορούν οι άνθρωποι να επικοινωνούν με μηχανήματα όπως συσκευές, ρομπότ, δεδομένα και υπηρεσίες; Επικεντρώνεται κυρίως στον σχεδιασμό, την αξιολόγηση και την εφαρμογή των τεχνολογιών επικοινωνίας και πληροφοριών, μεταξύ άλλων με συγκεκριμένο στόχο την ενίσχυση των πρακτικών των χρηστών, την εκτέλεση των καθηκόντων και την ποιότητα της εμπειρίας. Αυτή η αλληλεπίδραση έχει πρόσφατα αρχίσει να αναπτύσσεται και αναπτύσσεται με τη χρήση εφαρμογών TN και ETN. Η TN και η ETN έχουν ως στόχο την παροχή, ανάπτυξη, προώθηση και διερεύνηση νέων μεθοδολογιών και εμπειρικές γνώσεις σχετικά με αλληλεπιδραστικά γεγονότα ανθρώπου-μηχανής, που περιλαμβάνουν αλλά δεν περιορίζονται σε m-commerce, e-commerce, απασχόληση, εταιρείες, ανθρώπινες αλληλεπιδράσεις με έξυπνες μηχανές, τεχνολογίες, αλληλεπιδράσεις ανθρώπου-ρομπότ, καινοτόμες διαμορφώσεις διεπαφής για επαυξημένη πραγματικότητα, εικονική πραγματικότητα και ανάλυση προβλημάτων μέσα σε νευροφυσιολογικές συσκευές, μηχανές, και εργαλεία (π.χ., CT-Scan, EEG, MRI, GSR, και eye-trackers).

1.3. Προληπτική συντήρηση

Πρόοδος στις τεχνολογίες επικοινωνιών και πληροφοριών που μεταβάλλουν τις κατασκευαστικές βιομηχανίες για μια εξυπνότερη προσέγγιση, που επιτρέπει συνεχής και αυτόματη εξέταση των δεδομένων απόδοσης για την κατανόηση των εφαρμόσιμων διεισδύσεων που προβλέπουν τη δυσλειτουργία του εξοπλισμού, βελτίωση του χρόνου συνεχούς λειτουργίας και αύξηση της παραγωγικότητας μέσα από προγραμματισμένες και μη-χρονοβόρες διακοπές της παραγωγής για συντήρηση και επισκευή. Στην προληπτική συντήρηση, συλλέγονται πληροφορίες σε πραγματικό χρόνο για τον έλεγχο της κατάστασης των συσκευών. Ο στόχος είναι να ανακαλυφθούν πρότυπα που μπορούν να βοηθήσουν στην πρόβλεψη μέσα από την εφαρμογή μεθόδων TN για την επίτευξη αυτού του

στόχου. Όταν η προληπτική συντήρηση είναι μηχανογραφημένη, τα συστήματα καθίστανται πιο επιτακτικά κατά τον καθορισμό των απαιτήσεων των συσκευών και την πρόβλεψη της ολοκλήρωσης της συντήρησης. Ασφαλώς, η εφαρμογή των μεθόδων TN και ETN οδηγεί σε σημαντική μείωση του κόστους, μεγαλύτερη προβλεψιμότητα και μεγαλύτερη διαθεσιμότητα των συστημάτων.

1.4. Έξυπνα προϊόντα

Τα έξυπνα προϊόντα της Βιομηχανίας 4.0 χρησιμοποιούν IoT, υπολογιστική νέφους, κυβερνο-φυσικά συστήματα, μεθόδους τεχνητής νοημοσύνης βασιζόμενες σε ETN που επιτρέπουν στους ανθρώπους να αλληλοεπιδρούν με προϊόντα. Οι σημερινές μέθοδοι παραγωγής πρέπει να συνδυαστούν με τη Βιομηχανίας 4.0, η οποία επιτρέπει καινοτόμες τεχνολογίες για την πρόοδο των προϊόντων. Περιλαμβάνει προηγμένη, βελτιωμένη αυτόματη εμπειρία, γνώση και συστήματα παραγωγής εφαρμοσμένα σε πραγματικό χρόνο. Ως εκ τούτου, τα έξυπνα προϊόντα πρέπει να σχεδιάζονται με ισχυρές τεχνολογίες σε διασυνδεδεμένες ψηφιακές και περιβαλλοντικές μεθόδους.

1.5. Βιομηχανική Ρομποτική

Οι προηγμένες μέθοδοι TN και ETN επιτρέπουν στα βιομηχανικά ρομπότ να λειτουργούν και να ελέγχουν αυτόματα και διασυνδεδεμένα μεταξύ τους. Οι τυπικές εφαρμογές περιλαμβάνουν μηχανολογικές διαδικασίες όπως τη συγκόλληση, τη βαφή, αλλά και άλλες περίπλοκες διαδικασίες όπως την παλετοποίηση, την αποθήκευση και τη διερεύνηση προϊόντων [1]. Μέσα από τα ρομποτικά συστήματα όλα εκτελούνται με υψηλή αντοχή, ταχύτητα και ακρίβεια.

1.6. Ασφάλεια στον κυβερνοχώρο και ιδιωτικότητα

Ο συνδυασμός της τεχνολογίας των πληροφοριών και των επικοινωνιών αρχίζει με νέες προκλήσεις, κυρίως την ασφάλεια στον κυβερνοχώρο. Η ύπαρξη του IoT έχει επίσης αναμορφώσει σε μεγάλο βαθμό την παρουσία απειλών στον κυβερνοχώρο. Οι απειλές και τα τρωτά σημεία της IoT στον τομέα της ασφάλειας δημιουργούν τεχνικά εμπόδια, κύριες αιτίες για κυβερνοεπιθέσεις, απαιτήσεις κυβερνοασφάλειας και μεθόδους με παγκόσμια προοπτική, συμπεριλαμβανομένου το τμήμα τόσο ιδιωτικού όσο και δημόσιου τομέα. Η τεχνητή νοημοσύνη καθορίζει πτυχές ασφάλειας στον κυβερνοχώρο και προστασίας της ιδιωτικής ζωής, όπως η ανίχνευση απειλών στον κυβερνοχώρο σε έξυπνα συστήματα δικτύων της Βιομηχανίας 4.0. Συνήθως περιλαμβάνει την εξέταση της υποδομής του δικτύου και τον εντοπισμό απειλών για κυβερνοεπιθέσεις σε πραγματικό χρόνο. Επίσης συχνά περιλαμβάνει διερεύνηση κυκλοφορίας δικτύου, εντοπισμό τελικού σημείου, αναγνώριση και εντοπισμό λογισμικού κακόβουλης λειτουργίας. Τα συστήματα εντοπισμού απειλών στον κυβερνοχώρο που διαθέτουν δυνατότητες τεχνητής νοημοσύνης είναι συνήθως συστατικά στοιχεία μιας πιο ολοκληρωμένης ασφάλειας στον κυβερνοχώρο, λύση που χρησιμοποιεί πολλά προληπτικά μέτρα (π.χ., firewalls).

1.7. Έξυπνες Μεταφορές

Ο τομέας των μεταφορών βελτιώνεται επίσης ταχύτατα με την εμφάνιση της βιομηχανίας 4.0. Η έξυπνη τεχνολογία βοηθά στον καθορισμό του χρόνου μεταφοράς στο δρόμο ή δίνει συστάσεις για περαιτέρω αποτελεσματικά μέσα μεταφοράς, εναλλακτικά σχέδια διαδρομών, αναγνώριση σημάτων κυκλοφορίας, αυτόνομων ή αυτοκινούμενων οχημάτων. Επίσης βελτιώθηκαν οι μέθοδοι που βασίζονται στην τεχνητή νοημοσύνη, οι δυνατότητες μεταφοράς με τη χρήση προηγμένων συστημάτων που βασίζονται στην μηχανική όραση για την αυτοματοποιημένη ανίχνευση και παρακολούθηση οχημάτων. Το IoT, η υπολογιστική νέφους και η μεγάλη πρόοδος δεδομένων βελτιώνουν επίσης τις υπηρεσίες μεταφορών.

2. Μηχανική και Βαθιά Μάθηση

Η μηχανική μάθηση (Machine Learning ή εν συντομία ML) είναι μια υποομάδα της ΤΝ που περιγράφει μία από τις βασικές αρχές της ΤΝ. Μαθαίνει μέσα από εμπειρίες ή σύνολα δεδομένων και όχι μόνο από οδηγίες. Οι μέθοδοι που βασίζονται στην ML μαθαίνουν αυτόματα και βελτιώνουν τα αποτελέσματα του συστήματος μέσω εκπαίδευσης. Αυτές οι μέθοδοι εξετάζουν το τελικό αποτέλεσμα για κάθε αναγνωρίσιμο μοτίβο και αναζητούν πτυχές αντιστροφής μηχανικής για να παρέχουν ένα αποτέλεσμα. Αναπτύσσεται δηλαδή ένα σύστημα για τον τρόπο λήψης συμπερασμάτων και αποφάσεων με βάση προηγούμενες εμπειρίες που σκοπό έχει να μάθει μια συνάρτηση, η πιο συγκεκριμένα μια σχέση, ανάμεσα στα δεδομένα εισόδου και εξόδου. Υπάρχουν 4 μέθοδοι ώστε το μοντέλο να καταλάβει, ή να «μάθει», τη σχέση εισόδου-εξόδου και αυτές σύμφωνα με την βιβλιογραφία [8] είναι:

1. **Μάθηση με επίβλεψη:** Για κάθε είσοδο δίνεται η επιθυμητή έξοδος. Είναι η πιο κοινή μέθοδος που χρησιμοποιείται στην ML. Το αρνητικό αυτής της προσέγγισης είναι ότι χρειάζεται να είναι γνωστή η σχέση μεταξύ εισόδου-εξόδου, πρέπει δηλαδή να έχει οριστεί πριν από την εκπαίδευση, συνήθως με τη βοήθεια κάποιου ειδικού, αυτή η σχέση των δεδομένων εισόδου-εξόδου. Αυτό πολλές φορές δημιουργεί προβλήματα, κυρίως στον μεγάλο όγκο δεδομένων. Τα προβλήματα εμπίπτουν στο ότι χρειάζεται αρκετό ανθρώπινο δυναμικό, με ειδικευμένη γνώση και αρκετός χρόνος για να προετοιμαστούν τα δεδομένα και να δημιουργηθούν οι κανόνες, ή αλλιώς σχέσεις εισόδων-εξόδων.
2. **Μάθηση χωρίς επίβλεψη:** Το μοντέλο προσπαθεί να μάθει πρότυπα στα δεδομένα εισόδου χωρίς να γνωρίζει την επιθυμητή έξοδο. Τα προβλήματα δημιουργίας συστάδων (clustering) ανήκουν σε αυτή την κατηγορία. Με άλλα λόγια, αυτό το είδος μάθησης μπορεί να χρησιμοποιηθεί για την ομαδοποίηση πελατών βάσει των προτιμήσεών τους, για την ανάλυση κειμένου βάσει των θεμάτων που περιέχονται στο κείμενο, ή για τη μείωση του αριθμού των χαρακτηριστικών που χρησιμοποιούνται για την πρόβλεψη ενός φαινομένου. Αυτού του είδους μάθησης έχει απασχολήσει αρκετά τον κλάδο, καθώς μπορεί να δώσει λύσεις στον μεγάλο όγκο δεδομένων που δημιουργούνται καθημερινά, και που όπως προ-αναφέρθηκε, είναι δύσκολο να δημιουργηθούν ετικέτες, οι σχέσεις.
3. **Μάθηση με ημι-επίβλεψη:** Χρησιμοποιούνται κυρίως σε προβλήματα ταξινόμησης και ανίχνευσης ανωμαλιών [9]. Αυτό το είδος μάθησης εκμεταλλεύεται συνήθως την πληθώρα των δεδομένων μιας κλάσης, με σκοπό να κωδικοποιηθούν οι διαφορές μεταξύ των άλλων κλάσεων. Χρησιμοποιεί δηλαδή δεδομένα με και χωρίς ετικέτες ταυτόχρονα. Αυτό το είδος μάθησης είναι χρήσιμο όταν το κόστος, ή ο χρόνος για τη συλλογή και την επισήμανση των δεδομένων είναι υψηλό, αλλά υπάρχουν πολλά διαθέσιμα δεδομένα χωρίς ετικέτα.
4. **Μάθηση με ενίσχυση-επιβράβευση:** Η μάθηση με ενίσχυση βασίζεται στην ιδέα ότι ένας πράκτορας μπορεί να μάθει να επιλέγει βέλτιστες ενέργειες σε ένα περιβάλλον, λαμβάνοντας ανταμοιβή ή τιμωρία μέσα από τα αποτελέσματα των ενεργειών του. Η μάθηση με ενίσχυση είναι ένας πολύ δυναμικός και ευέλικτος τρόπος για να επιτύχουμε πολύπλοκους και δύσκολους στόχους, όπως η αυτόματη οδήγηση, η ρομποτική, κλπ. Η μάθηση με ενίσχυση απαιτεί τέσσερις βασικές συνιστώσες: έναν πράκτορα, ένα περιβάλλον, μια συνάρτηση ανταμοιβής και μια στρατηγική. Ο πράκτορας είναι το μοντέλο που μαθαίνει και λαμβάνει αποφάσεις. Το περιβάλλον είναι το σύνολο των καταστάσεων που αντιμετωπίζει ο πράκτορας. Η συνάρτηση ανταμοιβής είναι ένα κριτήριο που αξιολογεί τη συμπεριφορά του πράκτορα και του δίνει μια αριθμητική αξία ως αντάμειψη. Η στρατηγική είναι ο κανόνας που καθοδηγεί το πράκτορα να επιλέξει τη βέλτιστη ενέργεια σε κάθε κατάσταση.

Η βαθιά μάθηση [10] (Deep Learning ή εν συντομία DL) είναι ένα κομμάτι του ML που χρησιμοποιεί πολύ-επίπεδα τεχνητά νευρωνικά δίκτυα (ΤΝΔ) για να επιτυγχάνει πολύπλοκες και σύνθετες λειτουργίες, όπως για παράδειγμα:

- Η αναγνώριση εικόνων. Το κομμάτι αυτό είναι γνωστό και ως μηχανική όραση (Computer Vision) και είναι ένας τομέας έρευνας όπου παράγονται συστήματα που παρέχουν τη δυνατότητα σε υπολογιστές να "βλέπουν" να ερμηνεύουν και να κατανοούν ψηφιακά βίντεο και εικόνες. Στόχος του είναι να βγάλει συμπεράσματα από δεδομένα εικόνες και να τα χρησιμοποιήσει για να λύσει προβλήματα του πραγματικού κόσμου. Επί του παρόντος, υπάρχουν αρκετές εφαρμογές της μηχανικής όρασης στη βιομηχανία και οι προοπτικές είναι ακόμα μεγαλύτερες. Διάφοροι βιομηχανικοί οργανισμοί χρησιμοποιούν τη μηχανική όραση για τον εντοπισμό σφαλμάτων σε προϊόντα και εξοπλισμό, την αποφυγή εμποδίων σε κινήσεις αυτόνομων συστημάτων κ.α.
- Η σύνθεση κειμένου και η ανάλυση φωνής. Ο κλάδος αυτός είναι γνωστός ως επεξεργασία φυσικής γλώσσας (Natural Language Processing ή NLP). Η NLP είναι μια επιστήμη της κατανόησης, της διερμηνείας και της ανάγνωσης της γλώσσας από σύστημα ή μηχανήμα. Μόλις ένα σύστημα μπορεί να καταλάβει τι εννοεί ο χρήστης για να επικοινωνήσει, τότε το σύστημα αυτό έχει εφαρμογή. Όντας κομμάτι του DL, το NLP, είναι ουσιαστικά η ικανότητα ενός συστήματος ή μηχανής να μαθαίνει, να αντιλαμβάνεται και να κατανοεί την ανθρώπινη γλώσσα. Οι περισσότερες μέθοδοι NLP σήμερα χρησιμοποιούν τεχνικές που βασίζονται σε ΤΝΔ για να εξαγάγουν πληροφορίες από την ανθρώπινη γλώσσα.
- Η επεξεργασία δεδομένων χρονοσειρών. Τα δεδομένα χρονοσειρών είναι δεδομένα που συλλέγονται σε διαδοχικά χρονικά διαστήματα και απεικονίζουν την εξέλιξη μιας ή περισσότερων μεταβλητών με την πάροδο του χρόνου. Τα δεδομένα χρονοσειρών χρησιμοποιούνται για να αναλύσουν τις τάσεις, τους κύκλους, τη σταθερότητα και τη μελλοντική συμπεριφορά των μεταβλητών. Τα δεδομένα χρονοσειρών έχουν εφαρμογές σε πολλούς τομείς της βιομηχανίας όπως η παρακολούθηση διεργασιών, η απόδοση και η επίδοση των μηχανημάτων κ.α. Πρακτικά τα δεδομένα για την ανάλυση φωνής που αναφέρθηκε παραπάνω θεωρούνται και αυτά δεδομένα χρονοσειρών όποτε αρκετές μεθοδολογίες της προηγούμενης ενότητας έχουν εφαρμογή και σε πληθώρα προβλημάτων χρονοσειρών.

Η DL βασίζεται στην αρχή ότι ένα ΤΝΔ μπορεί να μάθει από τα δεδομένα που του δίδονται και να βελτιώσει την απόδοσή του μέσω της επανάληψης και της προσαρμογής. Ένα ΤΝΔ αποτελείται από πολλές μονάδες επεξεργασίας, που ονομάζονται νευρώνες, που συνδέονται μεταξύ τους μέσω συνάψεων. Κάθε νευρώνας λαμβάνει είσοδο από άλλους νευρώνες, υπολογίζει μια συνάρτηση ενεργοποίησης και παράγει μια έξοδο που στέλνεται σε άλλους νευρώνες. Τα νευρωνικά δίκτυα διαιρούνται σε επίπεδα (ή στρώματα), που μπορούν να είναι είσοδοι, έξοδοι ή κρυφά επίπεδα. Τα ΤΝΔ λειτουργούν σχεδόν με την ίδια αρχή όπως τα νευρωνικά κύτταρα του ανθρώπου. Μια ομάδα μεθόδων ή αλγορίθμων συσχετίζει διαφορετικές υποκείμενες μεταβλητές και αναλύει τα δεδομένα όπως κάνει ο ανθρώπινος εγκέφαλος. Η τεχνολογία DL εξελίσσεται με επιτυχία, καθώς τα συστήματα ή τα μοντέλα μπορούν να παρέχουν κορυφαία ακρίβεια. Υπάρχουν αρκετά είδη ΤΝΔ όπως για παράδειγμα τα επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks, ή RNN) [9], τα Συνελικτικά Νευρωνικά (Convolutional Neural Networks ή CNNs) [11], και τα γραφικά νευρωνικά δίκτυα (Graph Neural Networks ή GNN) [12]. Η DL περιέχει ΤΝΔ πολλών επιπέδων και πολλές φορές διαφορετικού είδους π.χ. CNN+RNN, και πρακτικά το βάθος των επιπέδων (ή πιο απλά το πλήθος των υπολογισμών παραμέτρων) μπορεί να φτάσει αριθμό πολλών εκατομμυρίων.

Ο μεγάλος αριθμός παραμέτρων στα Βαθιά Νευρωνικά Δίκτυα (Deep Neural Networks, ή εν συντομία DNNs) τους κάνει πολύπλοκους να κατανοηθούν και αναμφισβήτητα πιο δύσκολο να ερμηνευτούν. Ανεξάρτητα από την διασταυρούμενη ακρίβεια επικύρωσης ή άλλες παραμέτρους αξιολόγησης που θα μπορούσαν να υποδηλώνουν καλή μαθησιακή απόδοση, τα μοντέλα DL θα μπορούσαν εγγενώς να μάθουν ή να αποτύχουν να μάθουν αναπαραστάσεις από τα δεδομένα που ένας άνθρωπος θα μπορούσε να θεωρήσει σημαντικά. Για να εξηγηθούν οι αποφάσεις των DNN πρέπει να είναι γνωστές οι εσωτερικές λειτουργίες των DNNs. Ως εκ τούτου, συχνά η ικανότητα ερμηνείας των αποφάσεων τεχνητής νοημοσύνης θεωρείται δευτερεύουσα στον αγώνα για την επίτευξη έγκυρων αποτελεσμάτων ή για τη διασταύρωση της ακρίβειας σε ανθρώπινο επίπεδο. Επιπλέον η TN βασίζεται σε μεγάλα και περίπλοκα σύνολα δεδομένων, που μπορεί να περιέχουν αβεβαιότητες, ασάφειες, ασυμφωνίες ή προκατάληψη. Γι' αυτό, η TN μπορεί να παράγει αποτελέσματα που δεν αντικατοπτρίζουν την πραγματικότητα, τη λογική ή τη δικαιοσύνη.

3. Επεξηγηματική Τεχνητή Νοημοσύνη

Στην εργασία τους [13], ο Miller περιγράφει ότι η περιέργεια είναι ένας από τους κύριους λόγους για τους οποίους οι άνθρωποι ζητούν εξηγήσεις για συγκεκριμένες αποφάσεις. Ένας άλλος λόγος στην στροφή προς το ETN θα μπορούσε να είναι η διευκόλυνση της καλύτερης μάθησης - η επανάληψη του σχεδιασμού μοντέλου και η επίτευξη καλύτερων αποτελεσμάτων. Κάθε εξήγηση θα πρέπει να είναι συνεπής σε παρόμοια σημεία δεδομένων και να παρέχει σταθερή ή παρόμοια εξήγηση στο ίδιο χρονικό σημείο δεδομένων [14]. Οι εξηγήσεις πρέπει να κάνουν τον αλγόριθμο TN εκφραστικό για να βελτιώσουν την ανθρώπινη κατανόηση, την εμπιστοσύνη στη λήψη αποφάσεων και τη προώθηση αμερόληπτων και δίκαιων αποφάσεων. Έτσι, προκειμένου να διατηρηθεί η διαφάνεια, η εμπιστοσύνη και η δικαιοσύνη στη διαδικασία λήψης αποφάσεων της ML, απαιτείται μια εξήγηση ή μια ερμηνεύσιμη λύση για τα συστήματα ML.

Μια εξήγηση είναι ένας τρόπος ως προς την επαλήθευση της απόφασης εξόδου που λαμβάνεται από παράγοντα ή αλγόριθμο τεχνητής νοημοσύνης. Π.χ. για ένα μοντέλο ανίχνευσης σφαλμάτων σε προϊόντα που βγαίνουν από μια γραμμή παραγωγής, αν το μοντέλο χρησιμοποιεί εικόνες από τα παραγόμενα προϊόντα, μια εξήγηση μπορεί να είναι ένας χάρτης, ή μια περιοχή, των pixels που συμβάλλουν στην έξοδο του μοντέλου. Σε ένα άλλο παράδειγμα, για ένα μοντέλο αναγνώρισης ομιλίας, μια επεξήγηση της λειτουργίας του θα μπορούσε να είναι οι πληροφορίες του φάσματος ισχύος κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου, οι οποίες συνέβαλαν περισσότερο στην τρέχουσα απόφαση εξόδου. Οι εξηγήσεις μπορούν επίσης να βασίζονται στις παραμέτρους ή τις ενεργοποιήσεις των εκπαιδευμένων μοντέλων που εξηγούνται είτε με τη χρήση υποκατάστατων όπως τα δέντρα αποφάσεων ή με τη χρήση βαθμίδων ή άλλων μεθόδων. Στο πλαίσιο των αλγορίθμων ενισχυμένης μάθησης, μπορεί να δοθεί μια εξήγηση ως προς το γιατί ένας πράκτορας πήρε μια συγκεκριμένη απόφαση έναντι ενός άλλου. Ωστόσο, οι ορισμοί της ερμηνεύσιμης και εξηγήσιμης τεχνητής νοημοσύνης είναι συχνά γενικοί και ενδέχεται να είναι παραπλανητικοί. Μια συλλογή μοντέλων τεχνητής νοημοσύνης, όπως τα δέντρα αποφάσεων είναι εγγενώς ερμηνεύσιμα. Ωστόσο, επηρεάζονται από τα μειονεκτήματα του συμβιβασμού ερμηνευσιμότητας-ακρίβειας σε σύγκριση με τα μοντέλα DL [4].

Οι χρήσιμοι ορισμοί που δίνονται στην εργασία [4] ως προς τις έννοιες της ETN είναι:

- Ορισμός 1. Η **ερμηνευσιμότητα** είναι μια επιθυμητή ποιότητα ή χαρακτηριστικό ενός αλγορίθμου που παρέχει αρκετά εκφραστικά δεδομένα για να καταλάβουμε πώς λειτουργεί ο αλγόριθμος. Εδώ, οι ερμηνεύσιμοι τομείς θα μπορούσαν να περιλαμβάνουν εικόνες ή κείμενο που είναι

κατανοητό από τον άνθρωπο. Το λεξικό του Cambridge ορίζει: «Αν κάτι είναι ερμηνεύσιμο, είναι δυνατό να βρεθεί το νόημά του ή να βρεθεί ένα συγκεκριμένο νόημα σε αυτό».

Ορισμός 2. Η **ερμηνεία** είναι μια απλοποιημένη αναπαράσταση ενός σύνθετου τομέα, όπως τα αποτελέσματα που παράγονται από ένα μοντέλο μηχανικής μάθησης, σε ουσιαστικές έννοιες που είναι κατανοητές από τον άνθρωπο. Οι προβλέψεις εξόδου ενός απλού μοντέλου, όπως τα μοντέλα δέντρων αποφάσεων (decision trees), που βασίζονται σε κανόνες μπορούν εύκολα να ερμηνευτούν με διάσχιση του συνόλου κανόνων. Για παράδειγμα, ένα μοντέλο CNN θα μπορούσε να προσδιορίσει τα μέρη της εισόδου της εικόνας (pixels) που το οδήγησαν στην απόφαση.

Ορισμός 3. Μια εξήγηση είναι πρόσθετες **μετα-πληροφορίες**, που παράγονται από έναν εξωτερικό αλγόριθμο ή από το ίδιο το μοντέλο μηχανικής εκμάθησης, για να περιγράψουν τη σημασία ή τη σημασία ενός στιγμιότυπου εισόδου προς μια συγκεκριμένη έξοδο. Για μια εικόνα, ο χάρτης επεξήγησης μπορεί να είναι ένας χάρτης pixel με ίσο μέγεθος, ενώ για κείμενο μπορεί να είναι δείκτες με κάποιες τιμές επίδρασης ανά λέξη (score) που οδήγησαν στην απόφαση.

Ορισμός 4. Για ένα μοντέλο βαθιάς μάθησης f , αν οι παράμετροι του μοντέλου θ και οι πληροφορίες αρχιτεκτονικής του μοντέλου είναι γνωστές, το μοντέλο θεωρείται ότι λειτουργεί σαν με διαφάνεια, ή σαν «λευκό κουτί». Ένα τέτοιο μοντέλο κουτιού βελτιώνει τον εντοπισμό σφαλμάτων μοντέλου και προωθεί την εμπιστοσύνη. **Ωστόσο, γνωρίζοντας την αρχιτεκτονική του μοντέλου και τις παραμέτρους από μόνες τους δεν θα κάνει το μοντέλο εξηγήσιμο.**

Ορισμός 5. Ένα μοντέλο βαθιάς μάθησης f θεωρείται ως αδιάφανο μοντέλο, ή μοντέλο «μαύρο κουτί» αν οι παράμετροι του μοντέλου και οι αρχιτεκτονικές του δικτύου είναι **κρυμμένες από τον τελικό χρήστη**. Όπως συμβαίνει με την περίπτωση των νευρωνικών δικτύων όπου οι συνδέσεις των νευρώνων είναι άγνωστες προς τον χρήστη.

Σύμφωνα με την ίδια εργασία, η αναγκαιότητα για την ETN έχει τρεις πτυχές: 1) **την εμπιστευτικότητα**, 2) **την διαφάνεια**, και 3) **την αμεροληψία/δικαιοσύνη των αλγόριθμων** TN. Με τη χρήση των μη γραμμικών αλγορίθμων DL που αποτελούνται από εκατομμύρια παραμέτρους, οι τεχνικές ETN πρέπει να βελτιώσουν και τις τρεις πτυχές που αναφέρθηκαν παραπάνω. Πιο συγκεκριμένα η βελτίωση της διαφάνειας έχει ως σκοπό τη δημιουργία μιας κατανοητής από τον άνθρωπο αιτιολόγησης των αποφάσεων έτσι δίνεται ο επόμενος ορισμός:

Ορισμός 6. Ένα μοντέλο βαθιάς μάθησης θεωρείται διαφανές αν είναι αρκετά **εκφραστικό ώστε να είναι κατανοητό από τον άνθρωπο**. Εδώ, η διαφάνεια μπορεί να είναι μέρος της δομής του ίδιου του αλγορίθμου, ή να χρησιμοποιεί εξωτερικές τεχνικές, όπως αποσύνθεση του μοντέλου ή διάφορες προσομοιώσεις αποτελεσμάτων.

Η **διαφάνεια των μοντέλων** είναι σημαντική για την αξιολόγηση της ποιότητας των προβλέψεων. Ένα παράδειγμα δίνεται μέσω της [15]. Σε αυτό το πείραμα χρησιμοποιείται ένας αλγόριθμος με αντιμαχόμενα δίκτυα που είναι γνωστά στο ευρύ κοινό με την ορολογία adversarial networks. Εν συντομία αυτός ο αλγόριθμος χρησιμοποιεί δύο νευρωνικά δίκτυα που μαθαίνουν ταυτόχρονα και σκοπός του είναι το ένα δίνει καλύτερες αποφάσεις από το άλλο, περισσότερες πληροφορίες για τη λειτουργία μπορούν να δοθούν στην [15]. Σε αυτό το παράδειγμα κατηγοριοποίησης που παρουσιάζεται στην εικόνα, ένα εικονιζόμενο ζώο πάντα, κατηγοριοποιείται ως κλάση gibbon και μάλιστα με υψηλό ποσοστό σιγουριάς μετά την θελημένη προσθήκη θορύβου.

$$\begin{array}{ccc}
 \text{Image of a panda} & + .007 \times \text{Noisy image} & = \text{Modified image} \\
 x & \text{sign}(\nabla_x J(\theta, x, y)) & x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \\
 \text{"panda"} & \text{"nematode"} & \text{"gibbon"} \\
 57.7\% \text{ confidence} & 8.2\% \text{ confidence} & 99.3\% \text{ confidence}
 \end{array}$$

Εικόνα 3 Η αρχική εικόνα (x) που δείχνει ένα ζώο πάντα, δέχεται σκόπιμα θόρυβο στον αλγόριθμο αντιμαχόμενων δικτύων για να ξεγελάσει το δίκτυο που έχει ως στόχο να ταξινομήσει σωστά τις εικόνες. Με την εισαγωγή θορύβου το δίκτυο ξεγελιέται και δίνει μεγάλη σιγουριά (99.3%) ότι είναι εικόνα της κλάσης γίββων, αν και το αποτέλεσμα στο ανθρώπινο μάτι δεν έχει αλλάξει. Πηγή [15].

Η ουσία μέσα από την βελτίωση της **εμπιστευτικότητας** είναι να εμβαθύνει στην κοινωνική φύση του ανθρώπου. Η κοινωνική μας ζωή, οι αποφάσεις και οι κρίσεις μας βασίζονται κυρίως στη γνώση και τις διαθέσιμες εξηγήσεις για τις καταστάσεις και την εμπιστοσύνη που δημιουργούμε. Μια επιστημονική εξήγηση ή έστω μια λογική αιτιολόγηση κάτω του βέλτιστου για μια απόφαση, είναι αναμφίβολα καλύτερη από μια απόφαση με μεγάλη αυτοπεποίθηση χωρίς τις απαραίτητες εξηγήσεις. Άρα έτσι προκύπτει ο επόμενος ορισμός από τους ερευνητές ως:

Ορισμός 7. Η αξιοπιστία των μοντέλων DL είναι ένα μέτρο εμπιστοσύνης, των ανθρώπων που είναι εν τέλη οι τελικοί χρήστες, στην λειτουργία ενός μοντέλου στα προβλήματα του πραγματικού κόσμου, όπου αυτά τα προβλήματα θεωρούνται δυναμικά, δηλαδή αλλάζουν συνεχώς καταστάσεις και εξελίσσονται.

Συνεπώς, ο «λόγος για τον οποίο ελήφθη συγκεκριμένη απόφαση» είναι πρωταρχικής σημασίας για τη βελτίωση της εμπιστοσύνης των τελικών χρηστών, συμπεριλαμβανομένων των εμπειρογνομόνων, των προγραμματιστών, των νομοθετών και των μη ειδικών. Οι εξηγήσεις για την πρόβλεψη είναι πάντα τόσο σημαντικές για τους μετόχους και τις κυβερνητικές υπηρεσίες ώστε να θεμελιωθεί η σταθερότητα καθώς μεταβαίνουμε σε ένα συνδεδεμένο κοινωνικό-οικονομικό περιβάλλον που βασίζεται στην ΤΝ. Όσο αναφορά την βελτίωση σε **θέματα διακρίσεων, αμεροληψίας και δικαιοσύνης**, η ETN προάγει τη δικαιοσύνη και βοηθά στην άμβλυνση των προκαταλήψεων που εισάγονται στην απόφαση ΤΝ είτε από σύνολα δεδομένων εισόδου που δεν είναι κατάλληλα, αντιπροσωπευτικά και δίκαια, είτε από κακή αρχιτεκτονική-σχεδίαση νευρικού δικτύου. Έτσι προκύπτει ο όγδοος ορισμός:

Ορισμός 8. Η μεροληψία σε αλγόριθμους βαθιάς μάθησης δείχνει το δυσανάλογο βάρος, την προκατάληψη, την εύνοια ή κλίση του μαθησιακού μοντέλου προς υποσύνολα δεδομένων λόγω της μεροληψίας που προκλήθηκε με θεμιτούς ή αθέμιτους σκοπούς κατά τη συλλογή δεδομένων του ανθρώπου, ή λόγω των υπολειτουργιών που οφείλονται στον αλγόριθμο μάθησης.

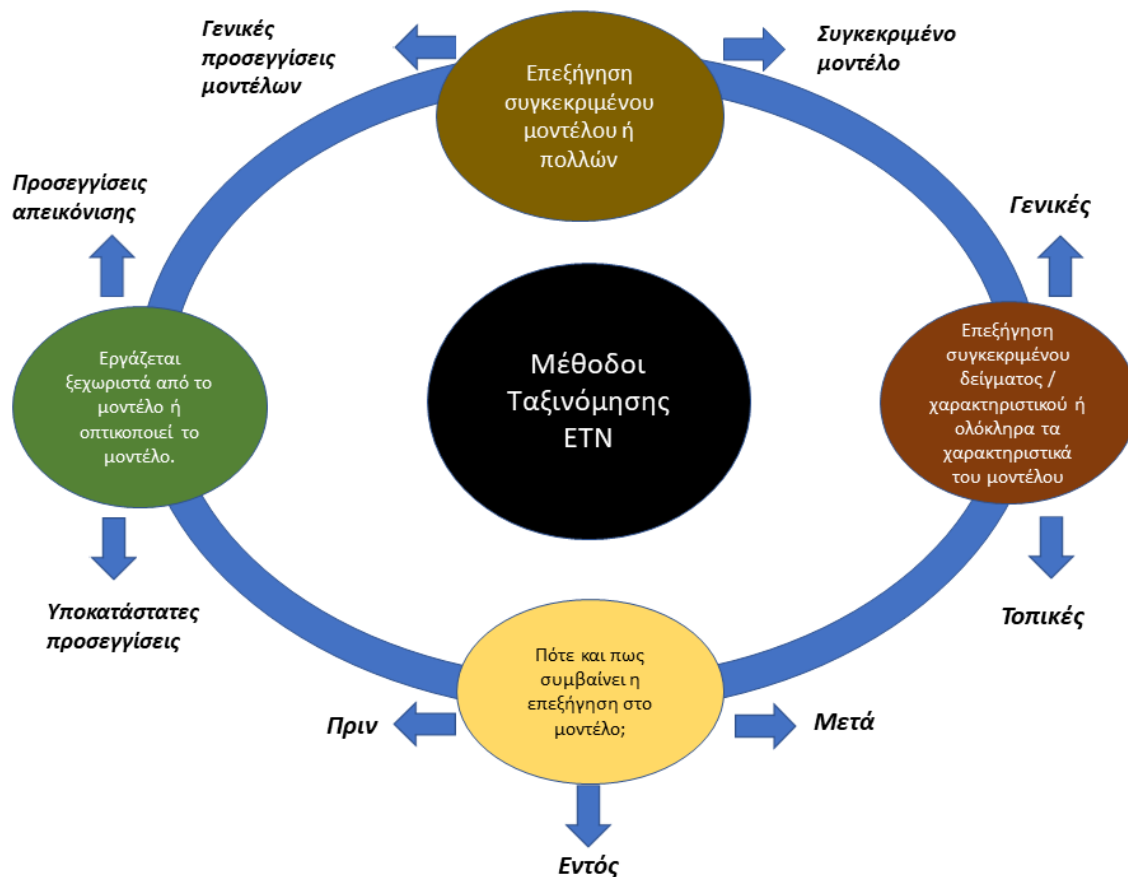
Η εκμάθηση της συμπεριφοράς του μοντέλου χρησιμοποιώντας τεχνικές ETN για διαφορετικές κατανομές δεδομένων εισόδου θα μπορούσε να βελτιώσει την κατανόηση της ασυμμετρίας και των προκαταλήψεων στα δεδομένα εισόδου. Αυτό θα μπορούσε να δημιουργήσει ένα ισχυρό μοντέλο ΤΝ. Βάση αυτού προκύπτει ο ένατος και τελευταίος ορισμός που δίνουν οι συγγραφείς:

Ορισμός 9. Η δικαιοσύνη στην DL είναι η ποιότητα ενός μαθησιακού μοντέλου στην παροχή αμερόληπτων και δίκαιων αποφάσεων χωρίς να ευνοείται κανένας πληθυσμός στην κατανομή των δεδομένων εισόδου.

Έχοντας θεσπίσει τους ορισμούς και την αναγκαιότητα της ETN, σειρά έχει να επεξηγηθεί ο τρόπος της εξήγησης που μπορεί να δοθεί μέσω της ETN. Παρακάτω παρουσιάζεται μια ταξινόμηση σύμφωνα με την [3].

3.1. Κατηγορίες Μεθόδων Επεξήγησης

Οι τεχνικές ETN ταξινομούνται σε 4 κατηγορίες ανάλογα την λειτουργία της τεχνικής, το πότε λαμβάνει μέρος και το ποιες λειτουργίες θέλει να εξηγήσει όπως φαίνεται και στο παρακάτω σχήμα (Εικόνα 4). Αυτές οι κατηγορίες αναλύονται περαιτέρω στις παρακάτω υπο-ενότητες.



Εικόνα 4 Οι 4 κατηγορίες ταξινόμησης μεθόδων ETN όπως ορίζονται στην [3].

Προσεγγίσεις σχετικές με ένα συγκεκριμένο μοντέλο και γενικές προσεγγίσεις μοντέλων (model agnostic and model specific).

Αυτή η κατηγορία μεθόδων περιλαμβάνει τις μεθόδους που ασχολούνται είτε να εξηγήσουν πολλά μοντέλα ή ένα συγκεκριμένο μοντέλο. Αυτές οι προσεγγίσεις στις εξηγήσεις παρουσιάζονται και εφαρμόζονται για την κατανόηση των αποφάσεων των μεθόδων ML και DL. Οι ειδικές προσεγγίσεις μοντέλου (model specific) εξαρτώνται από τις παραμέτρους των μοντέλων, π.χ., αλγόριθμοι που είναι γνωστοί στην διεθνή βιβλιογραφία ως:

- decision tree,
- Support Vector Machines,

- random forest,
- linear regression,
- extreme gradient boosting (XGBoost) για ταξινόμηση,
- επεξηγηματικά GNNs.

Στους αλγόριθμους που αναφέρθηκαν για την ερμηνευσιμότητα απαιτείται ακριβή γνώση των μοντέλων και των παραμέτρων τους ειδικά λόγω της περίπλοκης δομής των δεδομένων. Οι γενικές προσεγγίσεις που στην βιβλιογραφία ονομάζονται model agnostic, είναι κατά κύριο λόγο κατάλληλες για τις posthoc (μετά τη δημιουργία του μοντέλου) ερμηνείες και δεν περιορίζονται σε σταθερά συστήματα ή δομές μοντέλων. Οι προσεγγίσεις αυτές δεν παρέχουν άμεση πρόσβαση ή διαδρομή στα βάρη ή τις αρχιτεκτονικές παραμέτρους του κεντρικού μοντέλου, π.χ., οι τοπικές ερμηνεύσιμες αγνωστικιστικές εξηγήσεις μοντέλου (LIME) [16].

Η προσέγγιση **LIME**, εξηγεί το αποτέλεσμα ενός οποιαδήποτε μοντέλου ταξινόμησης με την μάθηση ενός άλλου διαφανούς μοντέλου που βασίζεται στις προβλέψεις του αδιάφανου μοντέλου για να μάθει. Η μέθοδος εξηγεί τον ταξινομητή για μια συγκεκριμένη μοναδική περίπτωση και είναι κατάλληλη για τοπικές εξηγήσεις, (όπως αυτές παρουσιάζονται στην επόμενη ενότητα). Με απλά λόγια η LIME χειρίζεται τα δεδομένα εισόδου και δημιουργεί μια σειρά από τεχνητά δεδομένα που περιέχουν μόνο ένα μέρος των αρχικών χαρακτηριστικών. Στη συνέχεια, χρησιμοποιεί ένα απλό μοντέλο, όπως ένα γραμμικό μοντέλο, για να προσεγγίσει τη συμπεριφορά του πρωτότυπου μοντέλου βάση των τεχνητών δεδομένων. Τέλος, επιλέγει τα χαρακτηριστικά που έχουν τη μεγαλύτερη επίδραση στην πρόβλεψη και τα παρουσιάζει ως εξήγηση.

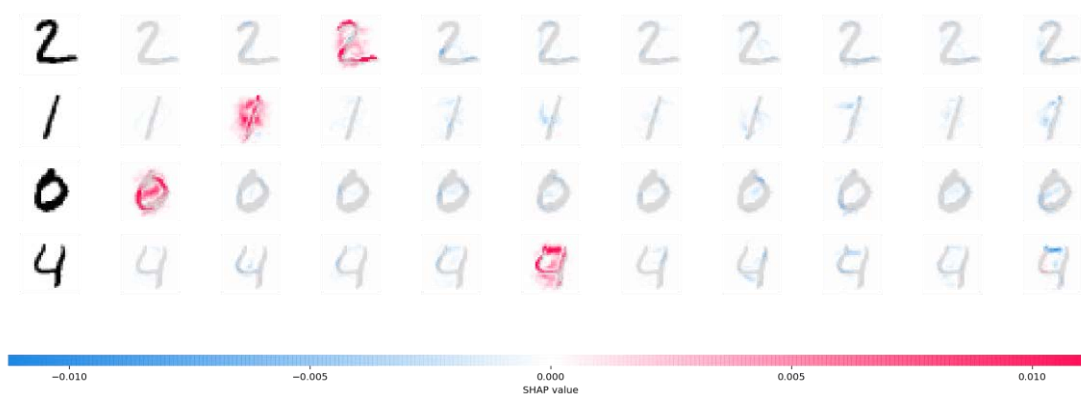
Τοπικές και γενικές προσεγγίσεις δεδομένων (local and global approaches)

Αυτή η κατηγορία ταξινομεί τις μεθόδους ανάλογα με το πώς επιχειρούν να δώσουν μια εξήγηση. Οι τοπικές ερμηνεύσιμες προσεγγίσεις είναι κατάλληλες για ένα συγκεκριμένο αποτέλεσμα του συστήματος. Μπορεί να επιτευχθεί με την περιγραφή μηχανισμών που καθορίζουν τον λόγο και το σκοπό για μια ειδική πρόβλεψη ή αποτέλεσμα. Για παράδειγμα, οι προγνώσεις του μοντέλου εγείρουν το ερώτημα «γιατί ένα μοντέλο ή ένα σύστημα για την συγκεκριμένη είσοδο έδωσε αυτή την έξοδο;». Για το λόγο αυτό αναπτύσσονται τοπικές ερμηνεύσιμες προσεγγίσεις. Γενικά, οι τοπικά επεξηγήσιμες μέθοδοι εστιάζουν σε μία μοναδική παρουσία δεδομένων εισόδου για να δημιουργήσουν εξηγήσεις χρησιμοποιώντας τα διαφορετικά χαρακτηριστικά των δεδομένων. Δηλαδή παρέχει εξηγήσεις σχετικά με τις ατομικές προβλέψεις. Σύμφωνα με την [4], στις τοπικές ερμηνεύσιμες προσεγγίσεις ανήκουν μεθοδολογίες όπως οι:

1. **Μεγιστοποίηση ενεργοποίησης (activation maximization).** Η αρχική έρευνα εδώ επικεντρώθηκε στο να καταλάβει τι μαθαίνει κάθε κρυφό στρώμα των νευρωνικών δικτύων και ειδικότερα των CNN. Σύμφωνα με τα αποτελέσματα τα αρχικά στρώματα μαθαίνουν υψηλού επιπέδου πληροφορίες (γωνίες, γραμμές, κτλπ.) των εικόνων. Όσο προχωράμε πιο βαθιά στα κρυφά στρώματα τότε τα CNN συνθέτουν πιο περίπλοκες πληροφορίες.
2. **Σημασιολογικοί χάρτες (Saliency Maps Visualization):** Αυτού του είδους οι μεθοδολογίες υπολογίζουν την κλίση (παράγωγο) για κάθε έξοδο των νευρωνικών δικτύων σύμφωνα με την είσοδο και έχουν εφαρμογή σε δεδομένα εικόνων. Αυτές οι προσεγγίσεις έχουν προσελκύσει αρκετό ενδιαφέρον μέσω μεθοδολογιών class activation mapping όπως το **Grad-CAM**[17]. Το Grad-CAM είναι ένας αλγόριθμος που χρησιμοποιείται για να εξηγήσει και να οπτικοποιήσει τη λειτουργία και τα αποτελέσματα ενός μοντέλου TN που βασίζεται σε CNN. Το Grad-CAM χρησιμοποιεί τις πληροφορίες των παραγώγων (gradients) που ρέουν στο τελευταίο επίπεδο συνέλιξης του μοντέλου για να δημιουργήσει έναν χάρτη ενεργοποίησης κλάσης (class

activation map), που αντικατοπτρίζει τις σημαντικές περιοχές στην εικόνα εισόδου για την πρόβλεψη μιας συγκεκριμένης κλάσης. Με αυτό τον τρόπο, μπορούμε να καταλάβουμε πώς το μοντέλο «βλέπει» και «επεξεργάζεται» μια εικόνα, και να ελέγξουμε αν η πρόβλεψή του είναι βάσει των σωστών χαρακτηριστικών ή αντικείμενων.

3. **Οπίσθια ανατροφοδότηση ανά στρώμα (layer-wise BackPropagation ή LRP).** Αυτή η μέθοδος υπολογίζει με τη μέθοδο backpropagation τιμές ενεργοποίησης, η αλλιώς σκορ, στα δεδομένα εισόδου αναλύοντας το μοντέλο σε κάθε στρώμα, από τα τελευταία προς το αρχικό στρώμα. Σε αυτή τη μέθοδο, κατά την εκπαίδευση ενός ΤΝΔ τα βάρη κάθε στρώματος υπολογίζονται και ανανεώνονται ξεχωριστά.
4. Η μεθοδολογία **LIME** που παρουσιάστηκε στην προηγούμενη υπο-ενότητα.
5. Η μεθοδολογία **SHAP** [18]. SHapley Additional exPlanations (SHAP) που βασίζονται στην θεωρία παιγνίων και σκοπό έχουν να δώσουν ένα σκορ στις εισόδους ανάλογα τη συνεισφορά τους στην απόφαση [18] λαμβάνοντας υπόψη όλες τις δυνατές συνδυασμούς των χαρακτηριστικών. Η μεθοδολογία SHAP έχει τραβήξει έντονο ενδιαφέρον γύρω της ειδικά για προβλήματα που βασίζονται σε δεδομένα μορφής πίνακα, δηλαδή με πολλές στήλες και εγγραφές. Με αυτό τον τρόπο, μπορούμε να καταλάβουμε πώς το μοντέλο αξιολογεί κάθε χαρακτηριστικό και να αποδώσουμε ευθύνη για την πρόβλεψή του. Για την εφαρμογή της μεθοδολογίας σε μοντέλα DL, οι ερευνητές ανέπτυξαν το Deep SHAP, ή αλλιώς βαθύ SHAP. Ένα παράδειγμα αυτής της μεθοδολογίας παρουσιάζεται στην Εικόνα 5 από την επίσημη ιστοσελίδα¹ της βιβλιοθήκης rpythοn που έχει υλοποιήσει και προσφέρει την μεθοδολογία αυτή.



Εικόνα 5 Παράδειγμα εφαρμογής του Deep SHAP στο γνωστό πακέτο δεδομένων αριθμών γραφικού χαρακτήρα MNIST. Με κόκκινο χρώμα οι περιοχές pixels που συνεισφέρουν σε μια κλάση ενώ με μπλε αυτές οι περιοχές που δεν συνεισφέρουν σε μια κλάση.

Αντίθετα, οι γενικές προσεγγίσεις εστιάζουν στον πυρήνα των συστημάτων και των μοντέλων, εφαρμόζοντας πλήρη γνώση των μοντέλων, μάθηση και συσχετισμένα δεδομένα. Επιδιώκεται να οριστεί η φύση, η συμπεριφορά και η απόδοση του μοντέλου ταυτόχρονα. Χαρακτηριστικής σημασίας είναι μια εξαιρετική απεικόνιση αυτής της προσέγγισης που διερευνά την εξεύρεση των χαρακτηριστικών που είναι γενικά υπεύθυνα ή υπόλογα για τις πιο αξιόπιστες επιδόσεις του μοντέλου μεταξύ όλων των διακριτών χαρακτηριστικών. Εξηγεί δηλαδή το πλήρες προγνωστικό μοντέλο στον χρήστη και τον βοηθά να ερμηνεύσει και να κατανοήσει την απόφαση. Η συμπεριφορά των μοντέλων ΤΝ για μια ακολουθία σημείων δεδομένων εισόδου θα μπορούσε να παρέχει πληροφορίες σχετικά με τα χαρακτηριστικά εισόδου, τα πρότυπα και τους συσχετισμούς εξόδου τους, προωθώντας έτσι τη διαφάνεια της συμπεριφοράς των μοντέλων. Διάφορες γενικά εξηγήσιμες μέθοδοι συνάγουν τα

¹ https://shap.lrbjball.readthedocs.io/en/latest/example_notebooks/deep_explainer/Front%20Page%20DeepExplainer%20MNIST%20Example.html

πολύπλοκα βαθιά μοντέλα σε γραμμικά αντίστοιχα που είναι πιο εύκολο να ερμηνευτούν. Τα πρότυπα που βασίζονται σε κανόνες και τα δενδρώδη μοντέλα, όπως τα δέντρα αποφάσεων, είναι εγγενώς καθολικά ερμηνεύσιμα. Η απόφαση εξόδου μεμονωμένων κλάδων ενός δέντρου μπορεί να ιχνογραφηθεί πίσω στην πηγή. Παρομοίως, τα γραμμικά μοντέλα συχνά εξηγούνται πλήρως με δεδομένες τις παραμέτρους του μοντέλου. Γενικά, οι γενικά εξηγήσιμες μέθοδοι δουλεύουν σε μια σειρά από εισόδους για να συνοψίσουν τη συνολική συμπεριφορά του αδιάφανου “black box” μοντέλου. Τέτοιες προσεγγίσεις είναι για παράδειγμα τα υποκατάστατα μοντέλα που αναλύονται παρακάτω και οι τοπικές μεθοδολογίες που μπορούν να τροποποιηθούν και να γίνουν γενικές προσεγγίσεις, τα νευρωνικά προσθετικά μοντέλα (**Neural Additive Models**) και η ανάλυση συσχέτισης φασματικής ταξινόμησης (**Spectral Relevance Analysis**) [4]. Μία διαδεδομένη γενική προσέγγιση είναι η σημαντικότητα μετάθεσης (**permutation importance**). Αυτή η τεχνική σκοπό έχει να δείξει ποια χαρακτηριστικά του πακέτου δεδομένων (στήλες) έχουν περισσότερο αντίκτυπο στην ακρίβεια και στο σφάλμα του μοντέλου. Έτσι σε κάθε πείραμα, το μοντέλο εκπαιδεύεται με ένα υποσύνολο των δεδομένων και η απόδοση του αναλύεται για να βρεθεί η συνεισφορά των συγκεκριμένων δεδομένων που ελήφθησαν.

Προσεγγίσεις πριν, εντός, και μετά την δημιουργία του μοντέλου (pro-model, in-model, post-model).

Αυτή η κατηγορία χωρίζει τις μεθόδους ανάλογα με το πότε ορίζονται οι επεξηγηματικές ιδιότητες του μοντέλου. Αν δηλαδή έχουν οριστεί πριν τη δημιουργία τους, κατά την λειτουργία του, ή αν είναι μετά την κατασκευή του. Οι προσεγγίσεις προ-μοντέλων (pro-model) είναι συνήθως αυτόνομες και δεν εξαρτώνται από ένα συγκεκριμένο μοντέλο ή αρχιτεκτονική συστήματος για την εφαρμογή του. Τα συστήματα ερμηνείας, τα οποία συνδυάζονται στο μοντέλο περιγράφονται ως προσεγγίσεις εντός του μοντέλου (in-model), όπως η εξόρυξη κανόνων σύνδεσης. Διαφορετικές προσεγγίσεις εκτελούνται αφού γίνει ένα σύστημα ή ένα μοντέλο, και ως εκ τούτου, αυτές οι τεχνικές ονομάζονται προσεγγίσεις μετά-του-μοντέλου (post-model). Οι προσεγγίσεις αυτές είναι ικανές να δημιουργήσουν πιθανώς σημαντικές διεισδύσεις σχετικά με το τι ακριβώς ένα σύστημα ή μοντέλο έμαθε κατά τη διάρκεια της εκπαίδευσης, όπως για παράδειγμα οι τιμές SHAP που αναφέρθηκαν.

Προσεγγίσεις απεικόνισης και υποκατάστατες προσεγγίσεις (visualization and surrogate).

Οι οπτικοποιήσεις δεν είναι ένα ξεχωριστό μοντέλο, αν και περιγράφουν κάποια στοιχεία των μοντέλων ή των συστημάτων από την οπτική αντίληψη, όπως π.χ. οι χάρτες ενεργοποίησης. Η βασική ιδέα είναι να κατανοήσουμε την ML και τα μοντέλα DL, ειδικά για να οπτικοποιήσουμε τις αναπαραστάσεις τους και για να εξερευνήσουμε και να αναλύσουμε τα κρυμμένα μοτίβα μέσα σε ένα νευρωνικό δίκτυο. Παρατηρείται ότι αυτές οι τεχνικές ανάλυσης δεν είναι περιοριστικές και βασίζονται σε διάφορες λογικές διαισθήσεις και, ως εκ τούτου, έχουν σημαντικές επικαλύψεις. Για παράδειγμα, ως επί το πλείστον τα post hoc μοντέλα (μοντέλα όπου αναλύονται μετά την δημιουργία τους), είναι model agnostic, καθώς αυτές οι προσεγγίσεις συχνά δεν έχουν πρόσβαση στην αρχιτεκτονική ενός συστήματος ή μοντέλου. Οι προσεγγίσεις υποκατάστασης περιλαμβάνουν διάφορα μοντέλα συλλογικά, και εφαρμόζονται για την αξιολόγηση διαφορετικών black box μοντέλων. Τα μοντέλα αυτά μπορούν να ερμηνευθούν περαιτέρω κατανοώντας τα συμπεράσματα του μοντέλου, εξετάζοντας τόσο την απόφαση του μοντέλου “black-box”, όσο και του υποκατάστατου μοντέλου.

3.2. Διαφανή μοντέλα και σύνοψη των Post-hoc τεχνικών ETN.

Σε μια αρκετά σημαντική και περιεκτική δουλειά που έγινε στην [19], συγκεντρώθηκαν όλες οι Post hoc τεχνικές για ETN. Ουσιαστικά όπως φαίνεται και στην Εικόνα 6 η ETN χωρίζεται σε δύο μεγάλες κατηγορίες, 1) στα διαφανή μοντέλα και 2) στις τεχνικές post-hoc που εφαρμόζονται στα μη διαφανή μοντέλα.

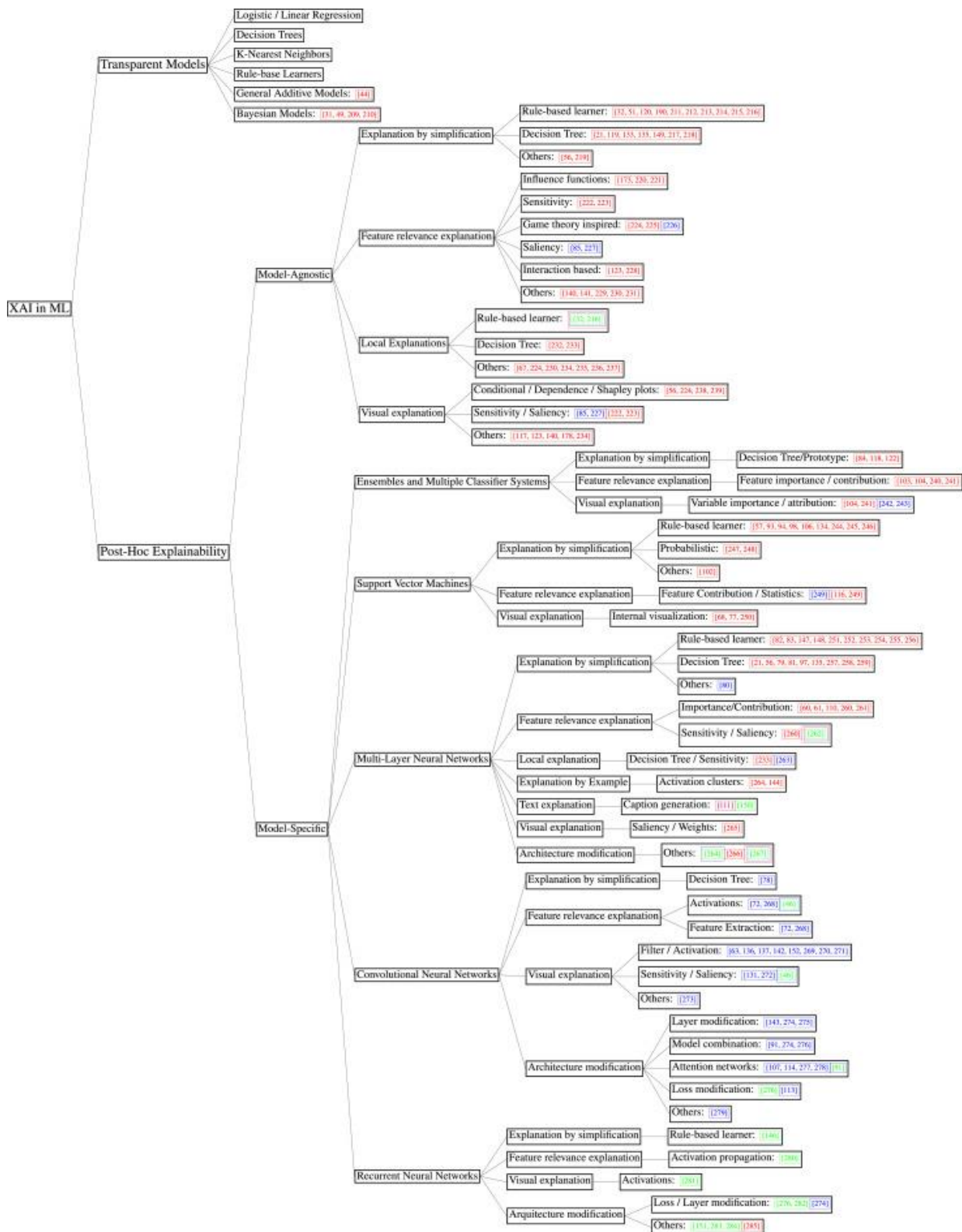
Τα διαφανή μοντέλα που παρουσιάζονται είναι:

- Linear logistic regression (γραμμική παλινδρόμηση). Είναι ένα μοντέλο ταξινόμησης για την πρόβλεψη μιας εξαρτημένης μεταβλητής (κατηγορίας) που είναι διχοτομική (δυναδική). Ωστόσο, όταν η εξαρτημένη μεταβλητή είναι συνεχής, η γραμμική παλινδρόμηση θα είναι το ομώνυμο της. Αυτό το μοντέλο λαμβάνει την υπόθεση της γραμμικής εξάρτησης μεταξύ των προγνωστικών και των προβλεπόμενων μεταβλητών, εμποδίζοντας μια ευέλικτη προσαρμογή στα δεδομένα. Ο συγκεκριμένος λόγος (ακαμψία του μοντέλου) είναι αυτός που διατηρεί το μοντέλο στην κατηγορία των διαφανών μεθόδων. Ωστόσο, η επεξήγηση συνδέεται με ένα συγκεκριμένο κοινό, γεγονός που κάνει ένα μοντέλο να εμπίπτει και στις δύο κατηγορίες ανάλογα με το ποιος θα το ερμηνεύσει. Με αυτόν τον τρόπο, η λογιστική και η γραμμική παλινδρόμηση, αν και πληρούν σαφώς τα χαρακτηριστικά των διαφανών μοντέλων (αλγοριθμική διαφάνεια, δυνατότητα αποσύνθεσης και προσομοίωση), μπορεί επίσης να απαιτούν τεχνικές μεταγενέστερης επεξήγησης (κυρίως οπτικοποίηση), ιδιαίτερα όταν το μοντέλο πρόκειται να εξηγηθεί σε μη εξειδικευμένο κοινό.
- Decision trees (δέντρα αποφάσεων). Τα δέντρα αποφάσεων είναι ένα άλλο παράδειγμα μοντέλου που μπορεί εύκολα να εκπληρώσει κάθε περιορισμό για διαφάνεια. Τα δέντρα αποφάσεων είναι ιεραρχικές δομές για τη λήψη αποφάσεων που χρησιμοποιούνται για την υποστήριξη προβλημάτων παλινδρόμησης και ταξινόμησης. Στην απλούστερη χρήση τους, τα δέντρα απόφασης είναι μοντέλα προσομοίωσης. Ωστόσο, οι ιδιότητές τους μπορούν να τα καταστήσουν αλγοριθμικά διαφανή. Η χρήση τους έχει συνδεθεί στενά με τα πλαίσια λήψης αποφάσεων, γεγονός που είναι ο λόγος για τον οποίο η πολυπλοκότητα και η κατανόησή τους θεωρούνταν πάντα πρωταρχικό ζήτημα. Ένα δέντρο προσομοίωσης αποφάσεων είναι αυτό που μπορεί να διαχειριστεί από έναν άνθρωπο χρήστη. Αυτό σημαίνει ότι το μέγεθός του είναι κάπως μικρό και η ποσότητα των χαρακτηριστικών και η σημασία τους είναι εύκολα κατανοητές. Μια αύξηση του μεγέθους μετατρέπει το μοντέλο σε αποσυνθέσιμο, καθώς το μέγεθός του εμποδίζει την πλήρη αξιολόγησή του (προσομοίωση) από άνθρωπο. Τέλος, η περαιτέρω αύξηση του μεγέθους του και η χρήση σύνθετων σχέσεων χαρακτηριστικών θα καταστήσει το μοντέλο αλγοριθμικά διαφανές χάνοντας τα προηγούμενα χαρακτηριστικά.
- K-Nearest Neighbors (K κοντινοί γείτονες). Μια άλλη μέθοδος που εμπίπτει σε διαφανή μοντέλα είναι αυτή του K-Nearest Neighbors (KNN), η οποία αντιμετωπίζει προβλήματα ταξινόμησης με έναν μεθοδολογικά απλό τρόπο: προβλέπει την κλάση ενός δείγματος δοκιμής ψηφίζοντας τις κλάσεις των K πλησιέστερων γειτόνων του (όπου η γειτονιά ή σχέση επάγεται από ένα μέτρο της απόστασης μεταξύ των δειγμάτων). Όταν χρησιμοποιείται στο πλαίσιο προβλημάτων παλινδρόμησης, η ψηφοφορία αντικαθίσταται από μια συγκέντρωση (π.χ. μέσος όρος) των τιμών-στόχων που σχετίζονται με τους πλησιέστερους γείτονες. Όσον αφορά την επεξήγηση του μοντέλου, είναι σημαντικό να παρατηρήσουμε ότι οι προβλέψεις που δημιουργούνται από τα μοντέλα KNN βασίζονται στην έννοια της απόστασης και της ομοιότητας μεταξύ των παραδειγμάτων, η οποία μπορεί να προσαρμοστεί ανάλογα με το συγκεκριμένο πρόβλημα που αντιμετωπίζεται. Είναι ενδιαφέρον ότι αυτή η προσέγγιση πρόβλεψης μοιάζει με αυτή της ανθρώπινης λήψης αποφάσεων που βασίζεται στην εμπειρία, η οποία αποφασίζει για το αποτέλεσμα παρόμοιων περιπτώσεων στο παρελθόν. Εκεί βρίσκεται η λογική του γιατί το KNN έχει επίσης υιοθετηθεί ευρέως σε περιβάλλοντα στα οποία η ερμηνευσιμότητα του μοντέλου είναι απαίτηση.
- Rule-based learning (μάθηση βάση κανόνων). Η μάθηση βάσει κανόνων αναφέρεται σε κάθε μοντέλο που δημιουργεί κανόνες για να χαρακτηρίσει τα δεδομένα από τα οποία προορίζεται να μάθει. Οι κανόνες μπορούν να λάβουν τη μορφή απλών κανόνων αν-τότε υπό όρους ή πιο περίπλοκων συνδυασμών απλών κανόνων για να σχηματίσουν τη γνώση τους. Επίσης,

συνδεδεμένα με αυτή τη γενική οικογένεια μοντέλων, είναι τα συστήματα που βασίζονται σε ασαφείς κανόνες, τα οποία είναι σχεδιασμένα για ένα ευρύτερο πεδίο δράσης, επιτρέποντας τον ορισμό προφορικά διατυπωμένων κανόνων σε ανακριβείς τομείς. Τα ασαφή συστήματα βελτιώνουν δύο βασικούς άξονες. Πρώτον, δημιουργούν πιο κατανοητά μοντέλα αφού λειτουργούν με γλωσσικούς όρους. Δεύτερον, αποδίδουν καλύτερα από τα κλασικά συστήματα κανόνων σε περιβάλλοντα με ορισμένους βαθμούς αβεβαιότητας-ασάφειας.

- General additive models (Γενικευμένα προσθετικά μοντέλα). Στις στατιστικές, ένα Γενικευμένο Προσθετικό Μοντέλο (GAM) είναι ένα γραμμικό μοντέλο στο οποίο η τιμή της μεταβλητής που πρόκειται να προβλεφθεί δίνεται από τη συνάθροιση ενός αριθμού άγνωστων ομαλών συναρτήσεων που ορίζονται για τις μεταβλητές πρόβλεψης. Ο σκοπός αυτού του μοντέλου είναι να συμπεράνει τις ομαλές συναρτήσεις των οποίων η αθροιστική σύνθεση προσεγγίζει την προβλεπόμενη μεταβλητή. Αυτή η δομή είναι εύκολα ερμηνεύσιμη, καθώς επιτρέπει στο χρήστη να επαληθεύσει τη σημασία κάθε μεταβλητής, δηλαδή, πώς επηρεάζει (μέσω της αντίστοιχης συνάρτησής της) την προβλεπόμενη έξοδο.
- Bayesian models. Ένα Bayesian μοντέλο συνήθως παίρνει τη μορφή ενός κατευθυνόμενου ακανόνιστου γραφικού μοντέλου του οποίου οι σύνδεσμοι αντιπροσωπεύουν τις υπό όρους εξαρτήσεις μεταξύ ενός συνόλου μεταβλητών. Για παράδειγμα, ένα Bayesian δίκτυο θα μπορούσε να αντιπροσωπεύει τις πιθανολογικές σχέσεις μεταξύ ασθενειών και συμπτωμάτων. Δεδομένων των συμπτωμάτων, το δίκτυο μπορεί να χρησιμοποιηθεί για τον υπολογισμό των πιθανοτήτων παρουσίας διαφόρων ασθενειών. Παρόμοια με τα GAM, αυτά τα μοντέλα μεταφέρουν επίσης μια σαφή αναπαράσταση των σχέσεων μεταξύ των χαρακτηριστικών και του στόχου, οι οποίες σε αυτή την περίπτωση δίνονται ρητά από τις συνδέσεις που συνδέουν τις μεταβλητές μεταξύ τους.

Όταν τα μοντέλα ML δεν πληρούν κανένα από τα κριτήρια που επιβάλλονται για να τα χαρακτηρίσουν διαφανή, πρέπει να επινοηθεί μια ξεχωριστή μέθοδος και να εφαρμοστεί στο μοντέλο για να εξηγηθούν οι αποφάσεις του. Αυτός είναι ο σκοπός των post-hoc τεχνικών επεξήγησης (γνωστές και ως επεξήγηση μετα-μοντελοποίησης), οι οποίες στοχεύουν στην επικοινωνία κατανοητών πληροφοριών σχετικά με τον τρόπο με τον οποίο ένα ήδη αναπτυγμένο μοντέλο παράγει τις προβλέψεις του για οποιαδήποτε δεδομένη είσοδο. Οι τεχνικές post-hoc ETN κατηγοριοποιούνται, κάνοντας διάκριση μεταξύ 1) εκείνων που έχουν σχεδιαστεί για την εφαρμογή τους σε μοντέλα ML οποιουδήποτε είδους (model agnostic). και 2) αυτά που έχουν σχεδιαστεί για ένα συγκεκριμένο μοντέλο ML και επομένως, δεν μπορούν να προεκταθούν άμεσα σε κανένα άλλο (model specific) όπως εξηγήθηκαν και στην προηγούμενη υπό ενότητα.



Εικόνα 6 Ταξινόμηση σύμφωνα με την [19] για τις τεχνικές ETN που βρέθηκαν στην βιβλιογραφία. Χρησιμοποιούν χρωματισμό στις παραπομπές που εμφανίζουν όπως μπλε, πράσινο και κόκκινο που αντιστοιχούν σε ETN τεχνικές που χρησιμοποιούν εικόνες, κείμενο ή δεδομένα σε μορφή πίνακα. Για την ταξινόμηση αυτή έψαξαν αναλυτικά στην βιβλιογραφία ώστε να συμπεριλάβουν μεθόδους που μπορούν να χρησιμοποιηθούν σε όλα τα μοντέλα.

Περαιτέρω πληροφορίες για την ταξινόμηση των τεχνικών που βρέθηκαν στην βιβλιογραφία παρουσιάζονται σύμφωνα με την [1], όπου συγκεντρώθηκαν 29 τεχνικές ETN βάση πολλαπλών

κριτηρίων και παρουσιάζονται στον Πίνακα 1. Οι περισσότερες τεχνικές ανήκουν στην κατηγορία των τοπικών επεξηγήσεων, ενώ 11 συνολικά τεχνικές βρέθηκαν για δεδομένα τύπου εικόνας.

Πίνακας 1 Ταξινόμηση τεχνικών ETN βάσει της [1] με ποικίλα κριτήρια.

Τεχνική	Model based	Attribution based	Example based	Τοπικές (L) / Γενικές (G)	Agnostic (A) / Specific (S)	Τύπος δεδομένων
All Convolutional Net	x	x		L	S	Εικόνα
Anchors		x		L/G	A	Κείμενο, πίνακας
Class Activation Mapping (CAM)		x		L	S	Εικόνα
Contextual Prediction Difference Analysis		x		L	S	Εικόνα
Deconvolutional Networks	x	x		L	S	Εικόνα
Deep Learning Important FeaTures (DeepLIFT)		x		L	S	Όλα
DICE			x	L	A	Όλα
DLIME	x	x		L	A	Όλα
GradCAM++		X		L	S	Εικόνα
Gradient		X		L	S	Όλα
Gradient * Input		X		L	S	Όλα
Gradient Weighted Class Activation Mapping (GradCAM)		X		L	S	Εικόνα
Integrated Gradients		X		L	S	Όλα
k-LIME	x	X		L	A	Όλα
Layer Wise Relevance Propagation (LRP)		X		L	A	Όλα
LIME	x	X		L	A	Όλα
LIMETree	x	X		L	A	Πίνακας
Local Foil Trees	x		x	L	A	Πίνακας
LoRE		X		L	A	Πίνακας
MAPLE	x	X		L	A	Πίνακας
Meaningfull Perturbation		X		L	S	Εικόνα
MMD-CRITIC			x	G	A	Όλα

Prediction Difference Analysis (PDA)		X		L	S	Εικόνα
RISE		X		L	S	Εικόνα
SHAP		X		L/G	A	Όλα
Smooth Grad		X		L	S	Εικόνα
STREAK				L	A	Εικόνα
TREPAN		x		G	S	Πίνακας

3.3. Μετρικά επίδοσης - αξιολόγησης

Η εξήγηση θεωρείται υποκειμενική έννοια. Ένα σύστημα τεχνητής νοημοσύνης μπορεί να εξηγηθεί είτε αν το μοντέλο είναι εγγενώς ερμηνεύσιμο είτε αν το μη ερμηνεύσιμο μοντέλο μπορεί να συμπληρωθεί με μια ερμηνεύσιμη και πιστή εξήγηση. Ενώ οι τεχνικές ETN παρέχουν διαφορετικά είδη πληροφοριών, η αντιληπτή ποιότητα των εξηγήσεων εξαρτάται από τους χρήστες, τον τομέα, την πληροφορία του ενδιαφέροντος, και την ίδια την εξήγηση. Για να αξιολογηθούν οι εξηγήσεις, είναι αναγκαίο να καθοριστούν διαφορετικά κριτήρια καταλληλότητας για μια εξήγηση. Δεδομένης μιας ερμηνεύσιμης προσέγγισης για μια αναφορά ένα μοντέλο απαριθμεί τέσσερις πτυχές που πρέπει να εξεταστούν στην αξιολόγηση: πιστότητα (ικανότητα για τη σωστή αποτύπωση της συμπεριφοράς του μοντέλου αναφοράς), σαφήνεια (ικανότητα παροχής μιας ενιαίας λογικής για την εξήγηση κάθε περίπτωσης δεδομένων), η δυνατότητα ερμηνείας (η προσέγγιση θα πρέπει να είναι κατανοητή από τον άνθρωπο) και η αλληλεπίδραση. Το σκέλος της αξιοπιστίας λαμβάνει υπόψη δύο ιδιότητες: την ορθότητα (το βαθμό στον οποίο κάθε συνιστώσα της επεξήγησης είναι πιστό στο μοντέλο αναφοράς) και την πληρότητα (το βαθμό στον οποίο η επεξήγηση περιγράφει την αναφορά). Απαριθμούνται άλλα τρία κριτήρια: ευαισθησία, ο βαθμός ολοκλήρωσης, και γνωστική υπεροχή. Η ευαισθησία ορίζεται ως η δύναμη της σχέσης των επεξηγηματικών μεταβλητών με τις συνθήκες του φόντου: όσο πιο αδύναμη είναι η σχέση, τόσο πιο πειστική είναι η εξήγηση. Ο βαθμός ολοκλήρωσης αναφέρεται στη σύνδεση της εξήγησης με ένα μεγαλύτερο θεωρητικό πλαίσιο. Τέλος, η γνωστική προβολή ορίζεται ως η ευκολία με την οποία μπορεί να ακολουθηθεί η λογική πίσω από την εξήγηση. Τα προαναφερθέντα κριτήρια απαιτούν διαφορετικές προσεγγίσεις αξιολόγησης. Προσδιορίστηκαν τρεις κατηγορίες από αυτές [1]:

- Αξιολόγηση εφαρμογής στήριξης: στηρίζεται σε μια πραγματική εφαρμογή, συλλέγει δεδομένα από ειδικούς σχετικά με τις εξηγήσεις που τους παρέχονται.
- Αξιολόγηση με βάση τον άνθρωπο: αναφέρεται στο feedback που λαμβάνεται από πειράματα που εκτελούν χρήστες, όταν δεν υπάρχει εφαρμογή πραγματικού κόσμου.
- Αξιολόγηση που βασίζεται στη λειτουργικότητα: η αξιολόγηση πραγματοποιείται με βάση κάποιο επίσημο ορισμό ή κριτήρια, που μετρά την ποιότητα της εξήγησης.

Για να αξιολογηθούν οι μέθοδοι, προτείνονται τρεις δοκιμές για αξιολογήσεις που βασίζονται στη λειτουργικότητα: Feature Augmentation Test, Synthetic Test και Feature Remove Test. Η δοκιμή αύξησης χαρακτηριστικών θεωρεί ότι αν οι τιμές των εξηγήσιμων χαρακτηριστικών από ένα συγκεκριμένο περιστατικό αντικαθίστανται από τις τιμές των χαρακτηριστικών αυτών από ένα παράδειγμα με διαφορετικό σήμα (π.χ., "new-label"), το αποτέλεσμα της ταξινόμησης θα πρέπει να είναι "new-label". Το Synthetic Test βασίζεται στην υπόθεση ότι αν τα χαρακτηριστικά εξηγησιμότητας επιλεγούν με ακρίβεια, νέες συνθετικές παρουσίες μπορούν να δημιουργηθούν διατηρώντας τις τιμές χαρακτηριστικού εξηγησιμότητας και εκχωρώντας τυχαίες τιμές στα υπόλοιπα χαρακτηριστικά χωρίς να επηρεαστεί το αποτέλεσμα πρόβλεψης. Τέλος, το κριτήριο της αφαίρεσης χαρακτηριστικών θεωρεί

ότι αν επιλεγούν σωστά τα επιλεγμένα χαρακτηριστικά εξηγησιμότητας, η αφαίρεση ενός από αυτά από την είσοδο θα πρέπει να οδηγήσει σε διαφορετική πρόβλεψη. Μολονότι η προσέγγιση αυτή υιοθετείται συχνά στη βιβλιογραφία, επισημαίνεται ότι στα δείγματα που αφαιρείται ένα υποσύνολο χαρακτηριστικών, έχουν διαφορετική κατανομή δεδομένων από ό,τι τα δείγματα στα οποία εκπαιδεύτηκε το μοντέλο, παραβιάζοντας μια βασική υπόθεση της μηχανικής μάθησης. Επί του παρόντος υπάρχει μικρή έρευνα σχετικά με την εφαρμογή και τις αξιολογήσεις με βάση τον άνθρωπο. Μια δημοφιλής και εξειδικευμένη μέθοδος είναι η αξιολόγηση για τη δημιουργία ενός heatmap σχετικά με την ευαισθησία του μοντέλου σε διαταραχές που βασίζονται στην περιοχή. Σύμφωνα με τον χάρτη θερμότητας, η κύρια ιδέα πίσω από αυτό είναι ότι η διαταραχή των σχετικών μεταβλητών εισόδου θα οδηγούσε σε μείωση της βαθμολογίας πρόβλεψης σε σύγκριση με τη διαταραχή των χαρακτηριστικών εισόδου με μικρότερη σημασία. Επίσης έχουν χρησιμοποιηθεί ερωτηματολόγια με σύντομες κλίμακες Likert. Αντίθετα, στην [20] έχουν χρησιμοποιήσει τρεις ποσοτικές μετρήσεις: ακρίβεια, χρόνο απόκρισης και υποκειμενική ικανοποίηση. Οι συντάκτες μέτρησαν την ακρίβεια και τον χρόνο απόκρισης σχετικά με την απόκριση του θέματος σε διάφορα καθήκοντα που προτείνονται στην έρευνά τους. Η υποκειμενική ικανοποίηση μετρήθηκε σε κλίμακα Likert για κάθε εξήγηση. Στην βιβλιογραφική ανασκόπηση [1], παρουσιάζονται 12 μαθηματικοί τύποι – μετρικά αξιολόγησης τεχνικών ETN βάση επικάλυψης κανόνων, ακρίβειας στις τοπικές εξηγήσεις, ποικιλία, διαφωνία κ.α.

$$HIS(x, R) = \begin{cases} 0, & \text{if } RT_{mean}(x, R) > RT_{max} \\ RT_{max} - RT_{mean}(x, M), & RT_{mean}(x, R) \leq RT_{max} \end{cases} \quad (1)$$

Εξίσωση 1: Βαθμολογία Ανθρώπινης Ερμηνευσιμότητας. Μετρά πόσο χρόνο χρειάζεται ο χρήστης για να προβλέψει την ετικέτα που έχει εκχωρηθεί σε ορισμένα σημεία δεδομένων, εκχωρώντας ένα όριο στον χρόνο απόκρισης. τα x και R αντιστοιχούν στην παρουσία και το μοντέλο.

$$I(x, y) = D_{KL}(P_{(x,y)} \parallel P_x \otimes P_y) \quad (2)$$

Εξίσωση 2: Αμοιβαία Πληροφόρηση. Μετρά την αμοιβαία εξάρτηση μεταξύ δύο τυχαίων μεταβλητών x και y. . Προτείνεται η μέτρηση της αμοιβαίας πληροφόρησης σε δύο περιπτώσεις: (i) μεταξύ των χαρακτηριστικών του αρχικού μοντέλου και του υποσυνόλου των εξηγήσιμων χαρακτηριστικών και (ii) έναντι των τιμών-στόχων. Ιδανικά, ο αριθμός των εξηγήσιμων χαρακτηριστικών θα πρέπει να μειωθεί για να μεγιστοποιηθεί η απλότητα και η ευρύτητα, ενώ παράλληλα θα πρέπει να επιδιώκεται η διατήρηση υψηλής πιστότητας όσον αφορά το σήμα-στόχο.

$$Diversity = \sum_{xi, xj \in E; xi \neq xj} \frac{d(xi, xj)}{2N_E} \quad (3)$$

Εξίσωση 3: Διαφορική μέτρηση. E είναι το σύνολο των παραδειγμάτων που λαμβάνονται υπόψη, d είναι ένα μετρικό απόστασης για τον χώρο X, ενώ το NE αντιστοιχεί στον αριθμό των παραδειγμάτων. Η πολυμορφία επιχειρεί να μετρήσει το βαθμό στον οποίο ένα σύνολο κανόνων ενσωματώνεται στην εξήγηση.

$$|A_0 \Delta X_0| \quad (4)$$

Εξίσωση 4: Μη ευαισθησία. Το A_0 αντιπροσωπεύει χαρακτηριστικά με μηδενική απόδοση, το X_0 αναφέρεται σε χαρακτηριστικά από τα οποία το μοντέλο δεν εξαρτάται λειτουργικά. $|\cdot|$ δηλώνει τον

καθορισμένο αριθμό στοιχείων συνόλου, και Δ η συμμετρική διαφορά συνόλου. Η διαίσθηση πίσω από τη μέτρηση μη ευαισθησίας είναι να εκτιμηθεί ότι η μέθοδος της δυνατότητας επεξήγησης δεν αποδίδει βαθμό συνάφειας στα χαρακτηριστικά από τα οποία δεν εξαρτάται το μοντέλο. Οι συγγραφείς την υπολογίζουν ως την συμμετρική διαφορά μεταξύ των χαρακτηριστικών που αποδίδονται με μηδενική απόδοση και των χαρακτηριστικών που το μοντέλο δεν εξαρτάται λειτουργικά. Αποτελεσματικά μετρά την πολυπλοκότητα αν κάποια χαρακτηριστικά εξήγησης μπορούν να αγνοηθούν χωρίς να επηρεάσουν σημαντικά την πρόβλεψη (βλ. 5).

$$k * = \underset{k \in i, \dots, N}{\operatorname{argmin}} |M_k| \text{ where } E(l(y *, f - M_k) | x * M_k) < \varepsilon \quad (5)$$

Εξίσωση 5: Αποτελεσματική Πολυπλοκότητα. Το M_k υποδηλώνει το σύνολο των κορυφαίων k χαρακτηριστικών, το x δηλώνει χαρακτηριστικά, $\varepsilon > 0$ αντιστοιχεί σε κάποια αυθαίρετη ανοχή, $f - M_k$ είναι ο περιορισμός του μοντέλου R σε μη σημαντικά χαρακτηριστικά, δεδομένου M_k .

$$\text{Disagreement}(R) = \sum_{i=1}^N |x| x \in D, x \text{ satisfies } q_i \wedge s_i, B(x) \neq c_i| \quad (6)$$

Εξίσωση 6: Μέτρηση διαφωνίας. Ποσοτικοποιεί τη διαφωνία μεταξύ ενός υποκατάστατου μοντέλου R και το μοντέλο πρόβλεψης αναφοράς B , με βάση ένα σύνολο δεδομένων D . Η τριάδα (q, s, c) αντιπροσωπεύει (χαρακτηριστικό, τελεστής, κατηγορία).

Προτείνονται άλλα έξι μετρικά για την αξιολόγηση των επεξηγήσεων πρόβλεψης: επικάλυψη κανόνα, κάλυψη, το μέγεθος του συνόλου κανόνων (βλ. 7), το μέγιστο πλάτος του συνόλου κανόνων, ο αριθμός των συνόλων περιγραφής και η επικάλυψη χαρακτηριστικών. Η επικάλυψη κανόνα υπολογίζει την επικάλυψη μεταξύ ζευγών κανόνων που ορίζονται στο υποκατάστατο μοντέλο. Αναμένεται ότι όσο χαμηλότερη είναι η επικάλυψη, τόσο χαμηλότερη είναι η ασάφεια του υποκατάστατου μοντέλου (βλ. 8). Η κάλυψη ορίζεται ως ο αριθμός των παρουσιών που ταιριάζουν με ένα δεδομένο κανόνα από το υποκατάστατο μοντέλο (δείτε, 9). Το μέγιστο πλάτος αναφέρεται στο μέγιστο πλάτος που προκύπτει από τον υπολογισμό του πλάτους σε όλα τα στοιχεία από το υποκατάστατο μοντέλο. Οι συντάκτες ορίζουν ένα στοιχείο είτε ως συνθήκες κανόνα είτε ως περιγραφές γειτονιάς (δείτε, 10). Οι συντάκτες ορίζουν τον αριθμό των μοναδικών συνόλων περιγραφής ως τον αριθμό των μοναδικών περιγραφών γειτονιάς που παρέχονται στο υποκατάστατο μοντέλο (βλ. 11) Τέλος, η επικάλυψη χαρακτηριστικών μετρά την επικάλυψη χαρακτηριστικών μεταξύ κάθε ζεύγους μοναδικής περιγραφής και κανόνα γειτονιάς (βλέπε, 12).

$$\text{RuleSetSize}(R) = \text{Number of Rules}(q, s, c) \quad (7)$$

Εξίσωση 7: Μέγεθος συνόλου κανόνων. Το R υποδηλώνει το σύνολο αποφάσεων. Η τριάδα (q, s, c) αντιπροσωπεύει (χαρακτηριστικό, χειριστής, κατηγορία). Η τριάδα περιέχεται στο σύνολο των αποφάσεων.

$$\text{RuleOverlap}(R) = \sum_{i=1}^N \sum_{j=1, j \neq i}^N \text{overlap}(q_i \wedge s_i, q_j \wedge s_j) \quad (8)$$

Εξίσωση 8: Επικάλυψη κανόνα. Υποδηλώνει το σύνολο αποφάσεων. Η τριάδα (q, s, c) αντιπροσωπεύει (χαρακτηριστικό, χειριστής, τιμή).

$$cover(R) = |x|x \in D, xsatisfies q_i \wedge s_i, where i \in 1 \dots N| \quad (9)$$

Εξίσωση 9: Κάλυψη. Το R υποδηλώνει το σύνολο αποφάσεων. Η τριάδα (q, s, c) συμβολίζει (χαρακτηριστικό, τελεστής, τιμή). Το D αντιπροσωπεύει ένα σύνολο δεδομένων, και το x την παρουσία σε ένα τέτοιο σύνολο δεδομένων.

$$MaximumWidth(R) = \max(width(e)), e \in \bigcup_{i=1}^N (q_i \cup s_i) \quad (10)$$

Εξίσωση 10: Μέγιστο Πλάτος. Το R υποδηλώνει το σύνολο αποφάσεων. Το e αντιπροσωπεύει στοιχεία, τα οποία μπορεί να είναι είτε συνθήκες κανόνα ή περιγραφές γειτονιάς.

$$Number\ Of\ Unique\ Descriptor\ Sets(R) = |dset(R)|, where\ dset(R) = \bigcup_{i=1}^N (q_i) \quad (11)$$

Εξίσωση 11: Αριθμός Μοναδικών Περιγραφών Συνόλων. Το R δηλώνει το σύνολο αποφάσεων και το q δηλώνει χαρακτηριστικά.

$$FeatureOverlap(R) = \sum_{i=1}^N FeatureOverlap(q, s_i) \quad (12)$$

Εξίσωση 12: Επικάλυψη των χαρακτηριστικών. Το R δηλώνει το σετ των αποφάσεων, το q δηλώνει τα χαρακτηριστικά στα σετ του επεξηγητή, και το s δηλώνει χειριστές.

4. Μέθοδοι Επεξηγηματικών Τεχνικών που χρησιμοποιήθηκαν στη βιομηχανία

Οι τεχνολογίες που σχετίζονται με την Βιομηχανία 4.0 και είναι ενσωματωμένες σε προηγμένες μεθόδους που βασίζονται στην TN και την ETN επιτυγχάνουν αξιοσημείωτη επιτυχία, ακρίβεια και ποιότητα σε μια πληθώρα εφαρμογών. Εξηγήσιμες και ερμηνεύσιμες αρχιτεκτονικές μπορούν να επιτύχουν ένα μεγάλο επίπεδο κατανόησης κερδίζοντας έτσι μεγάλη προσοχή σε όλους τους τομείς της Βιομηχανίας 4.0. Μέσα από την βιβλιογραφία έγινε η συλλογή των ευρημάτων για τεχνικές ETN που εφαρμόστηκαν σε πραγματικά βιομηχανικά προβλήματα. Τα ευρήματα παρουσιάζονται σε αυτή την ενότητα.

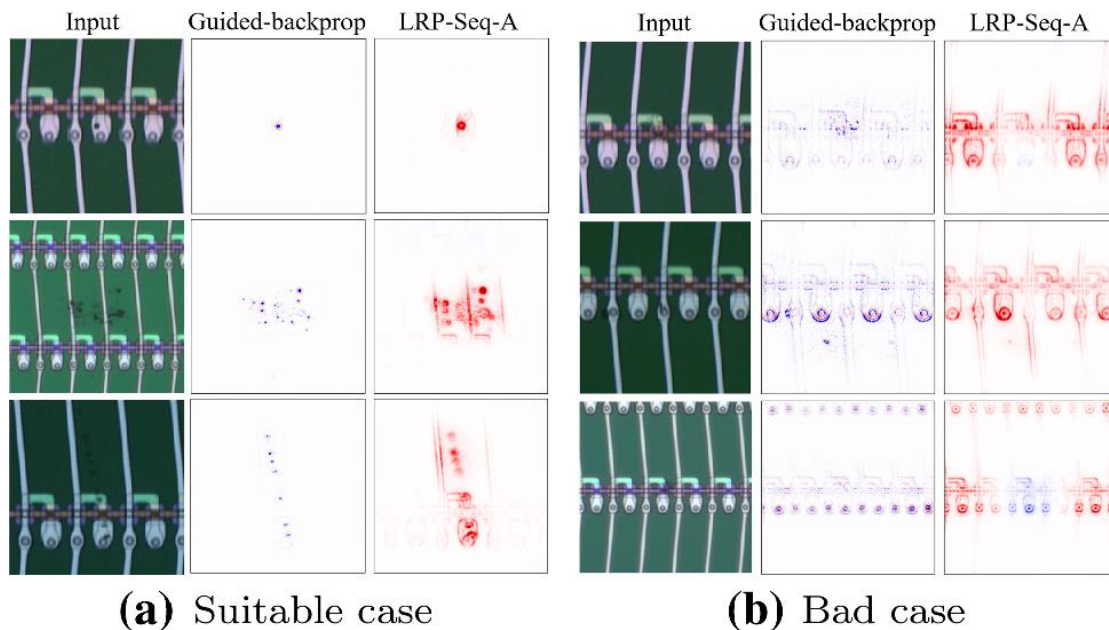
Αν και υπάρχουν πολλαπλές μέθοδοι ETN, δεν επαρκούν από μόνες τους για να παρέχουν κατανοητές εξηγήσεις στους τελικούς χρήστες. Για το λόγο αυτό είναι ενσωματωμένες σε πλαίσια και εφαρμογές που παρέχουν ένα βολικό περιβάλλον διεπαφής και πρόσθετα πλαίσια για την επίτευξη αυτού του στόχου κατανόησης. Στην εργασία [1] έγινε μια συλλογή από τέτοιες πλατφόρμες-πλαίσια-εργαλεία τα οποία παρουσιάζονται εδώ. Για παράδειγμα, ένα τέτοιο πλαίσιο είναι το bLIMEy [21], το οποίο αποσυνθέτει τα υποκατάστατα μοντέλα σε τρία βήματα: 1) ερμηνεύσιμη αναπαράσταση δεδομένων (μετασχηματισμός δεδομένων από το αρχικό στο ερμηνεύσιμο πεδίο), 2) δειγματοληψία δεδομένων και 3) παραγωγή εξηγήσεων. Μια άλλη παρόμοια προσέγγιση που επιχειρεί να υλοποιήσει ένα γενικό εργαλείο εξήγησης παρουσιάστηκε στο [22] και περιγράφει το σύστημα που ονόμασαν IBEX (Interactive Black-box EXplanation system). Το IBEX είναι ένα πλαίσιο με δύο λειτουργίες: 1) έναν εξηγητή που παρέχει εξηγήσεις του μοντέλου με βάση τις ανάγκες του χρήστη, και 2) ένα στοιχείο δειγματοληψίας, που επιλέγει κατάλληλα στοιχεία για να δημιουργήσει την εξήγηση. Το AI Explainability 360 που παρουσιάστηκε στην [23] είναι μια επεκτάσιμη εργαλειοθήκη ανοιχτού

λογισμικού, που έχει αναπτυχθεί και παρέχει επεξηγηματικές λειτουργίες, μετρικά και μια εύχρηστη πλατφόρμα λογισμικού κατά το στάδιο ανάπτυξης και λειτουργίας του μοντέλου AI, κατά το είδος του μοντέλου και τις απαιτήσεις επεξήγησης που έχουν οριστεί. Στην δουλειά που παρουσιάστηκε στην [24], γίνεται εξερεύνηση της χρησιμότητας των σημασιολογικών τεχνολογιών σε αφηρημένες σχετικές έννοιες που βρίσκονται κωδικοποιημένες στις λειτουργίες του μοντέλου. Αυτό έχει ως αποτέλεσμα να αποφεύγεται η έκθεση ευαίσθητων λεπτομερειών-πληροφοριών σχετικά με το μοντέλο πρόβλεψης και ταυτόχρονα να παρέχει πληροφορίες υψηλότερου επιπέδου που θα είναι πιο ευκολά κατανοητές στους χρήστες. Οι συγγραφείς συμπληρώνουν τις επεξηγήσεις των μοντέλων με πληροφορίες σχετικά με πραγματικά γεγονότα που αναφέρθηκαν στα μέσα ενημέρωσης και πιθανόν επηρέασαν τις μεταβλητές ενδιαφέροντος Συλλέγει δηλαδή τα σχόλια του χρήστη για μια δεδομένη πρόβλεψη, τις επεξηγήσεις, τις προτεινόμενες αποφάσεις και κινήσεις του χρήστη. Παρόμοιες λειτουργίες αναπτύχθηκαν και στην επόμενη εργασία που βρέθηκε [25] που συλλέγει γνώση από τους χρήστες. Ανέπτυξαν έναν ευφυή βοηθό για την βιομηχανία, ο οποίος δημιουργεί επεξηγηματικές οδηγίες για τους χρήστες χρησιμοποιώντας μια αυτοδίδαχτη μέθοδο (heuristic) και γνώση που είναι διαθέσιμη στον κλάδο. Η εφαρμογή ουσιαστικά παρακολουθεί την έμμεση και ρητή ανατροφοδότηση του χρήστη σχετικά με τοπικές εξηγήσεις πρόβλεψης, επιτρέποντας αξιολογήσεις βασισμένες στην εφαρμογή.

Η ενσωμάτωση των μεθόδων επεξηγησιμότητας σε εφαρμογές επιτρέπει την παροχή σχετικών πληροφοριών σχετικά με τις προγνώσεις μοντέλων σε διάφορους ενδιαφερόμενους. Για παράδειγμα, οι επιστήμονες που ασχολούνται με την επεξεργασία δεδομένων και οι μηχανικοί που εφαρμόζουν ML και DL χρησιμοποιούν δεδομένα «χαμηλού επιπέδου» ή ανεπεξέργαστα δεδομένα, για την παρακολούθηση της συμπεριφοράς του μοντέλου TN. Μέσα από αυτή τη δομή των δεδομένων εντοπίζουν μη αποδεκτές περιπτώσεις λειτουργίας και εργάζονται προς την βελτίωση μέσα από ένα πιο ακριβές και ισχυρό μοντέλο. Από την άλλη πλευρά, οι εργαζόμενοι και οι επόπτες απαιτούν «υψηλού επιπέδου» πληροφορίες, δηλαδή ιδέες η έννοιες, που φανερώνουν τους λόγους πίσω από τις προβλέψεις του μοντέλου. Έτσι μπορούν να διερευνήσουν αλληλεπιδραστικά διαφορετικά σενάρια "what-if" και να παρέχουν ανατροφοδότηση σχετικά με τις παρεχόμενες εξηγήσεις. Άρα οι μέθοδοι επεξηγησιμότητας μπορούν να είναι χρήσιμες σε ένα ευρύ φάσμα των περιπτώσεων που απαιτούνται στην έξυπνη βιομηχανία, όπως η αυτόματη ανίχνευση ελαττωμάτων (ενημερώνει τον χρήστη για τις περιοχές της εικόνας που επηρεάζουν την απόφαση), σχεδιασμό παραγωγής (παρέχει πληροφορίες για το κόστος της ευκαιρίας, δίνοντας διαφορετικές αποφάσεις προγραμματισμού παραγωγής), ή πρόβλεψη της ζήτησης (παρέχει πληροφορίες στο γιατί περιμένουμε ότι η ζήτηση θα λάβει χώρα και ποιοι παράγοντες επηρεάζουν τις εκτιμήσεις ποσότητας).

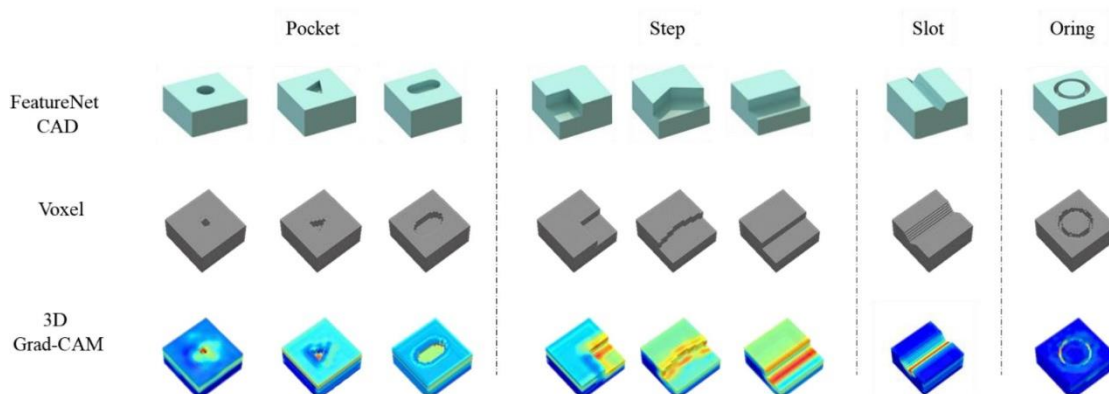
Αρκετές τεχνικές επεξηγησιμότητας έχουν εφαρμοστεί στον τομέα της βιομηχανίας και συγκεκριμένα στον τομέα της προληπτικής διαχείρισης ποιότητας (Quality 4.0) για την ενίσχυση της διαφάνειας των μοντέλων TN. Για παράδειγμα στην [26], χρησιμοποιήθηκαν τεχνικές ETN, όπως CAM και αντιθετικοί σημασιολογικοί χάρτες με βάση τον υπολογισμό της παράγωγου (gradient-based), για να εξηγηθούν τα αδιαφανή μοντέλα ταξινόμησης. Εφαρμόστηκε στον τομέα που ασχολείται με την ποιότητα των συγκολλήσεων, και πιο συγκεκριμένα στη υπερηχητικής συγκόλλησης μεταλλικών πόλων μπαταριών. Παρήγαγαν θερμούς χάρτες (heatmaps) με τους οποίους οπτικοποίησαν αρκετούς χάρτες με χρώμα για να καταλάβουν την συμπεριφορά του μοντέλου για τις σωστές προβλέψεις (True Positives) και για τις λανθασμένες προβλέψεις (False Positives). Στην επόμενη εργασία [27] εφαρμόστηκαν διάφορες μέθοδοι ETN για να παρέχουν επεξηγήσεις της ταξινόμησης ελαττωμάτων. Εφαρμόστηκε στον τομέα που ασχολείται με πάνελ οθόνης με τεχνολογία υγρών κρυστάλλων λεπτής μεμβράνης-τρανζίστορ. Τεχνικές όπως οι CAM, LRP, ενσωματωμένες παράγωγοι, καθοδηγούμενη αντίστροφη μετάδοση και SmoothGrad υλοποιήθηκαν και απεικονίστηκαν σε ένα μοντέλο ταξινόμησης VGG-16. Με βάση τα αποτελέσματα οπτικοποίησης, επιλέχθηκαν οι τεχνικές LRP και καθοδηγούμενη αντίστροφη μετάδοση, καθώς παρήγαγαν καλά κατανοημένους θερμούς χάρτες. Ένα τέτοιο παράδειγμα από την εργασία

τους φαίνεται στην Εικόνα 7. Επιπλέον, με την τοποθέτηση του μοντέλου σε έναν αλγόριθμο δέντρου αποφάσεων και τη μετατροπή των αποτελεσμάτων πρόβλεψης σε ανθρώπινο ερμηνεύσιμο κείμενο, οι συγγραφείς πέτυχαν το μέγιστο επίπεδο επεξηγησιμότητας. Αυτό το συμπέρασμα προέκυψε όταν παρουσίασαν τα αποτελέσματα σε ειδικούς του τομέα για σκοπούς αξιολόγησης.



Εικόνα 7 Παράδειγμα της οπτικοποίησης των αποτελεσμάτων των μεθόδων που προτάθηκαν στην [27]. Αριστερά (a) φαίνεται ξεκάθαρα η περιοχή που το μοντέλο αναγνώρισε με ελάττωμα, ενώ δεξιά (b) κάποιες περιοχές που λανθασμένα ενεργοποιήθηκαν.

Στον τομέα εκτίμησης του κόστους κατασκευής περιγράφεται μια μέθοδος που βασίζεται στην απεικόνιση των χαρακτηριστικών κατεργασίας ενός μοντέλου σχεδιασμού 3D με τη βοήθεια υπολογιστή που επηρεάζουν την αύξηση του κόστους κατασκευής [28]. Για το προτεινόμενο σκοπό αυτό, εφαρμόστηκε η μεθοδολογία 3D Grad-CAM ως μέθοδος ETN. Πιο συγκεκριμένα, η προσέγγιση τους βασίζεται στην συλλογή δεδομένων, στην εφαρμογή μοντέλων DL 3D και στην οπτικοποίηση για την επεξήγηση των αποτελεσμάτων. Με την χρήση του 3D Grad-CAM μπόρεσαν να δείξουν ότι το μοντέλο καταλαβαίνει τις CNC μηχανικές λειτουργίες και να διαφοροποιήσει την δυσκολία τους, όπως φαίνεται στην Εικόνα 8.

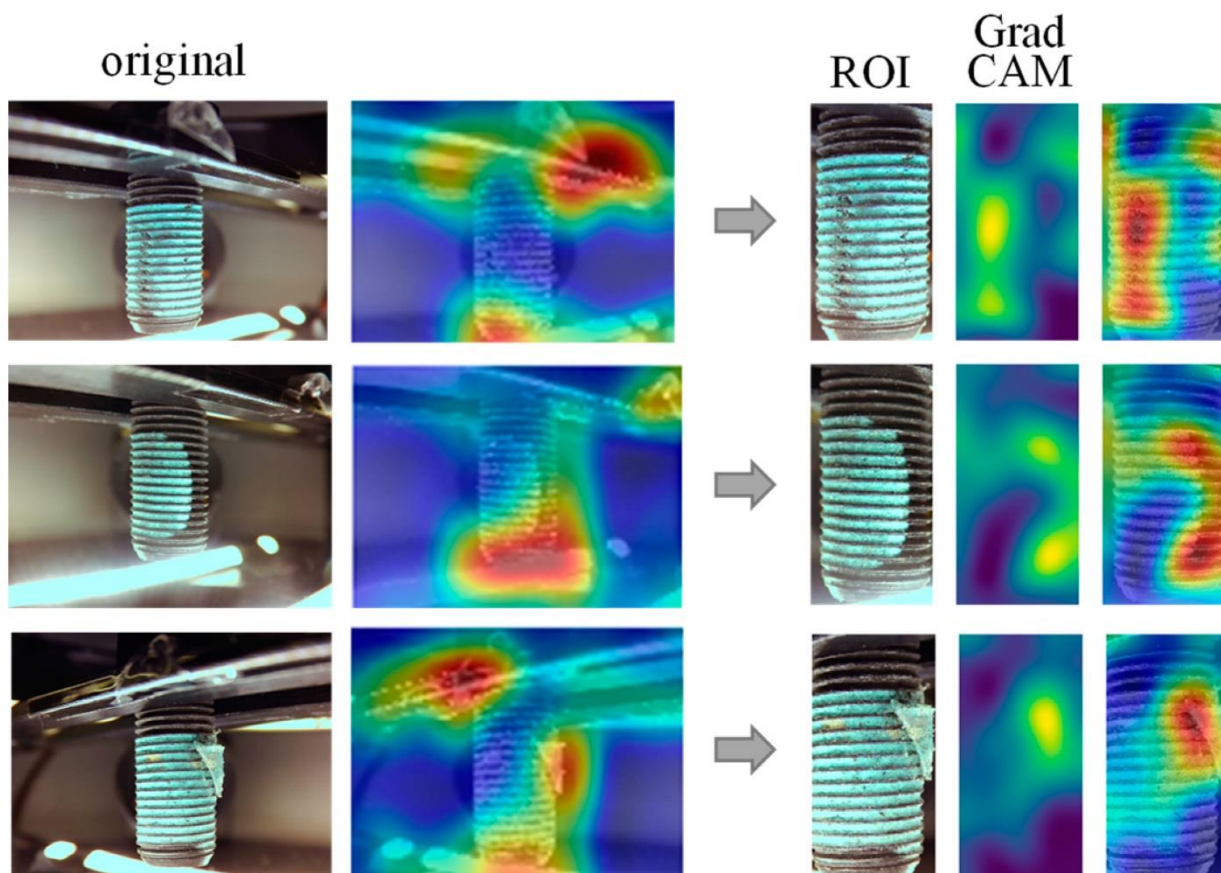


Εικόνα 8 Η εφαρμογή της μεθοδολογίας 3D Grad-CAM που φανερώνει της περιοχές με δυσκολία κατασκευής [28].

Στην [29] οι ερευνητές παρουσίασαν μια εφαρμογή ETN στα πλαίσια της Βιομηχανίας 4.0. Πιο συγκεκριμένα εφάρμοσαν μια μεθοδολογία πρόβλεψης βασισμένη στην DL στο εργοστάσιο DFKI-

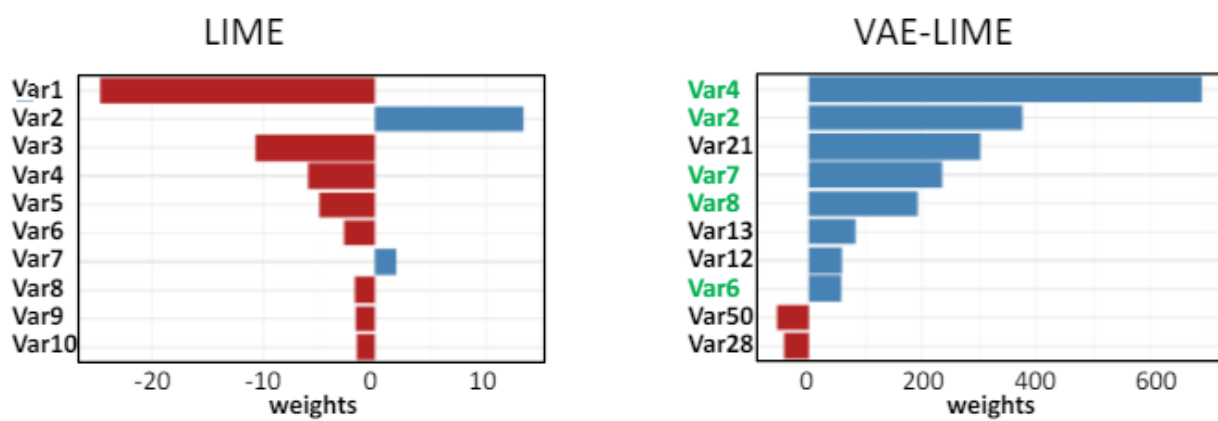
Smart-Lego-Factory. Το μοντέλο τους, που χρησιμοποιεί επαναλαμβανόμενα νευρωνικά δίκτυα μπορεί να προβλέψει με ακρίβεια το αποτέλεσμα μιας διεργασίας λαμβάνοντας δεδομένα της παραγωγής. Για να κάνουν το μοντέλο τους διαφανή προς τους εργαζόμενους και τους διευθυντές του εργοστασίου ώστε να μπορούν να το καταλάβουν, τεκμηριώσουν και να προσαρμόσουν την γραμμή ανάλογα οι ερευνητές πρόσθεσαν επεξηγηματικές ιδιότητες post-hoc (μετα-ανάλυση ή μετά-του-μοντέλου). Χρησιμοποίησαν τοπικές εξηγήσεις αλλά και γενικές εξηγήσεις. Πιο συγκεκριμένα για την γενική εξήγηση χρησιμοποίησαν μια προσέγγιση που δίνει τιμή συνεισφοράς σε κάθε ομάδα δεδομένων (feature) που βασίζεται στα βάρη σύνδεσης. Για τοπικές εξηγήσεις χρησιμοποίησαν κανόνες και την μεθοδολογία λογιστικής παλινδρόμησης για να δημιουργήσουν ένα διαφανές μοντέλο που θα δίνει πληροφορίες συνεισφοράς των εκάστοτε εισόδων.

Χάρτες θερμότητας Grad-CAM χρησιμοποιήθηκαν στην [30] για να κάνουν ποιοτικό έλεγχο στις βίδες και συγκεκριμένα στο υγρό επικάλυψης που χρησιμοποιείται για να αποφευχθεί η χαλάρωση τους. Οι ερευνητές χρησιμοποίησαν DL και CNNs και βρήκαν ότι ένα μοντέλο VGG-16 με μεταφορά γνώσης ήταν το πιο αποδοτικό για να βρει τα ελαττωματικά προϊόντα. Ένα πακέτο από υγιή και ελαττωματικά προϊόντα δημιουργήθηκε μέσα από 4 κάμερες που έπαιρναν φωτογραφίες από όλες τις πλευρές των βιδών. Η ακρίβεια του VGG-16 έφτασε στο 97% και μέσα από τα αποτελέσματα Grad – CAM έβγαλαν χρήσιμα συμπεράσματα για την λειτουργία του αλγόριθμου. Συγκεκριμένα παρατήρησαν ότι εκτός από την επικάλυψη ο αλγόριθμος είχε ευαισθησία και σε άλλους εξωτερικούς παράγοντες που φαινόταν στις εικόνες. Για το λόγο αυτό , έγινε περικοπή στα σημεία ενδιαφέροντος στις εικόνες (Regions Of Interest) ώστε να βελτιωθεί η απόδοση του μοντέλου. Τα αποτελέσματα ETN Grad -CAM φαίνονται στην παρακάτω εικόνα.



Εικόνα 9 Τα αποτελέσματα της ETN μεθόδου Grad-CAM που χρησιμοποίησαν στην [30].

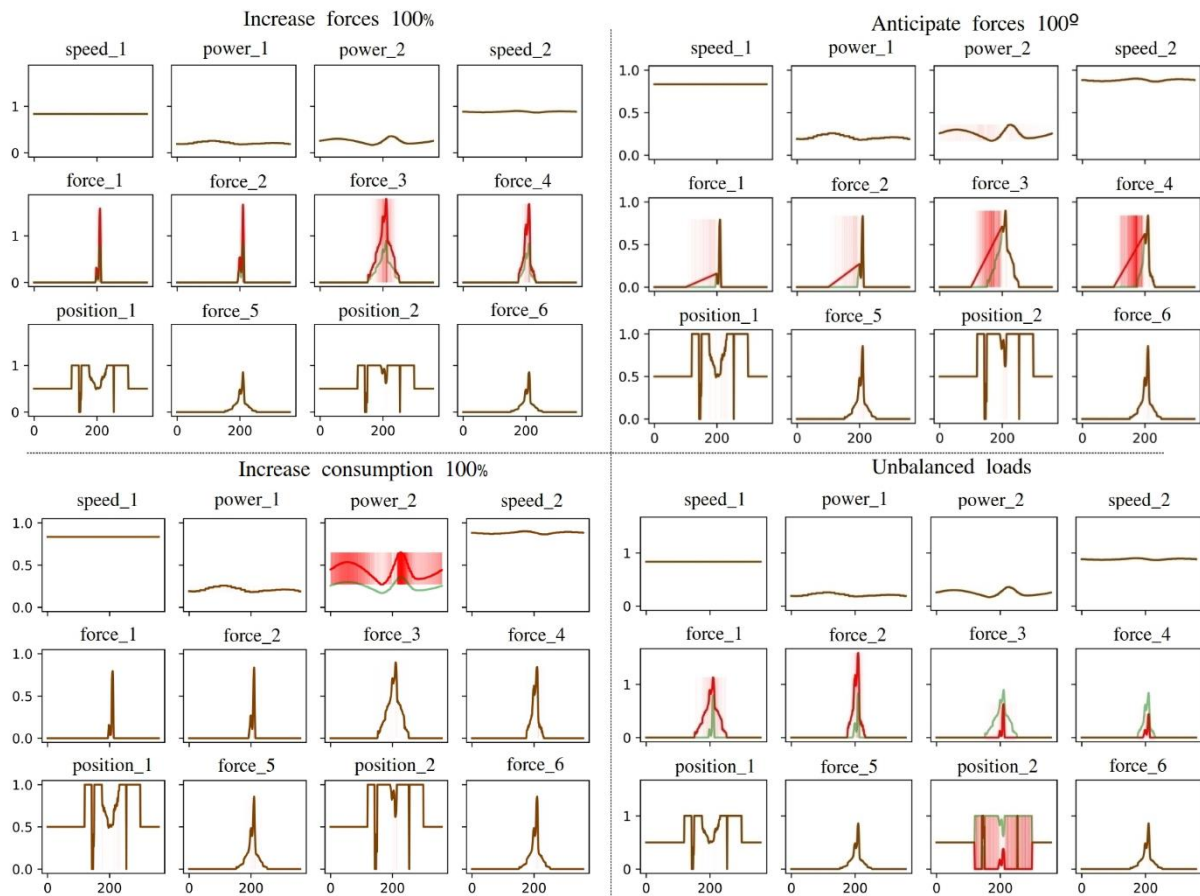
Στην εργασία [31], παρουσιάζεται μια νέα προσέγγιση, η VAE-LIME, για την τοπική ερμηνεία μοντέλων που βασίζονται σε δεδομένα και προβλέπουν τη θερμοκρασία του υγρού μετάλλου που παράγεται από έναν υψικάμινο. Τα δεδομένα αυτής της διαδικασίας παραγωγής σιδήρου χαρακτηρίζονται από πολυμεταβλητές χρονοσειρές με υψηλό αλληλοσυσχέτιση, που αντιπροσωπεύουν την υποκείμενη διεργασία στον υψικάμινο. Η συνεισφορά μας έγκειται στη χρήση ενός μετατροποποιητή διακωδικοποίησης με μεταβλητότητα (VAE) για να μάθει τα σύνθετα χαρακτηριστικά της διαδικασίας του υψικαμίνου από τα δεδομένα. Ο στόχος του VAE είναι να δημιουργήσει βέλτιστα τεχνητά δείγματα για την εκπαίδευση ενός τοπικά ερμηνεύσιμου μοντέλου που θα αντιπροσωπεύει καλύτερα το "black-box" μοντέλο στη γειτονιά του δείγματος εισόδου που επεξεργάζεται το "black-box" μοντέλο για να κάνει μια πρόβλεψη. Σε σύγκριση με τη LIME, η VAE-LIME παρουσιάζει σημαντικά βελτιωμένη τοπική πιστότητα του τοπικά ερμηνεύσιμου γραμμικού μοντέλου με το "black-box" μοντέλο, οδηγώντας σε μια πιο ισχυρή ερμηνεία του μοντέλου. Η Εικόνα 10 συγκρίνει τις δέκα πιο σημαντικές μεταβλητές που δίνονται από τις μεθόδους VAE-LIME και LIME.



Εικόνα 10 Σύγκριση της τεχνικής LIME με την VAE-LIME για διαφορές μεταβλητές της εργασίας [31].

Εφαρμογές ETN στην πρόβλεψη της συντήρησης εξετάστηκαν στην εργασία [32]. Οι ερευνητές χρησιμοποίησαν δεδομένα όπως καταγεγραμμένα σφάλματα στις μηχανές κατά την παραγωγή, δεδομένα από τηλεμετρία, δεδομένα συντήρησης και άλλα δεδομένα από τις μηχανές παραγωγής για να αναπτύξουν μεθόδους ETN σε 5 datasets. Προ-επεξεργάστηκαν κατάλληλα τα δεδομένα ώστε να τα συμπτύξουν και να μπορέσουν να τα χρησιμοποιηθούν για την ανάπτυξη μοντέλων. Το μοντέλο ML που εξετάστηκε είναι ένα Gradient Boosting Machine για ταξινόμηση πολλών κλάσεων. Αυτό το μοντέλο χρησιμοποιεί αλγόριθμο μάθησης με βάση το δέντρο απόφασης που όπως αναφέρθηκε πιο πάνω στην βιβλιογραφία ταξινομείται ως διαφανή μοντέλο. Στο ίδιο τομέα, της πρόβλεψης συντήρησης, βρέθηκε στην βιβλιογραφία ακόμα μια λεπτομερής δουλειά [33]. Σε αυτή την εργασία συγκεντρώθηκαν δεδομένα από μια μηχανή πρέσας. Αρχικά χρησιμοποιήθηκαν μόνο υγιή δεδομένα για να αναπτυχθούν και να εκπαιδευτούν μοντέλα εντοπισμού σφαλμάτων σε δεδομένα χρονοσειρών. Διάφοροι δημοφιλή αλγόριθμοι εντοπισμού σφαλμάτων χρησιμοποιήθηκαν όπως δύο διαστάσεων autoencoders, null-space, Principal Component Analysis, support vector machines και extreme learning machine. Επίσης συνδύασαν τεχνικές t-SNE, gaussian mixture models, SOM και OPTICS για να ενσωματώσουν μια κατανόηση στον τρόπο λειτουργίας και μάθησης των τεχνικών που χρησιμοποιήθηκαν. Για να αναπτύξουν τις επεξηγηματικές ιδιότητες των μοντέλων χρησιμοποιήθηκε το GradientExplainer που συνδυάζει integrated Gradients, SHAP τιμές και SmoothGrad σε μία συνάρτηση. Τα αποτελέσματα της ETN δείχνουν, ότι ο αλγόριθμος διάγνωσης μπορούν να ανιχνεύσουν ποια σήματα προκαλούν την ανωμαλία με μια προσέγγιση πολλών μεταβλητών και έχοντας χρησιμοποιήσει μάθηση με ημι-επίβλεψη, που σημαίνει ότι δεν είδαν τέτοια περιστατικά κατά την εκπαίδευση. Η παρακάτω εικόνα δείχνει τα αρχικά δεδομένα ενός πρεσαρίσματος με πράσινο χρώμα, με κόκκινο τα επεξεργασμένα δεδομένα που έχουν ένα είδος δυσλειτουργίας, και το υπόβαθρο της εικόνας έχει μια κόκκινη

επισκίαση του μετρικού εξήγησης. Δείχνει δηλαδή σε ποια σημεία του κύκλου λειτουργίας τα δεδομένα παρουσιάζουν ανωμαλίες.



Εικόνα 11 Διάγνωση με την χρήση ETN για την ίδια καταγραφή πρεσαρίσματος με 4 τύπους δυσλειτουργίας, σκιαγραφίζοντας με κόκκινο χρώμα τα σημεία όπου δημιουργούν τις δυσλειτουργίες.

Πακέτα λογισμικού

Η έρευνα [4] παραθέτει μερικά εργαλεία ανοιχτού κώδικα (open-source), που ανακάλυψαν στην πλατφόρμα κώδικα github στον τομέα της ETN. ανοιχτού κώδικα έχουν βελτιώσει σημαντικά την αναπαραγωγή στην έρευνα και έχουν αποτελέσει πραγματικό δώρο για την εξέλιξη της έρευνας στους κλάδους της TN, της DL, αλλά και της ETN.

- Πακέτο Interpret. Διατίθεται από την TranslateML και μπορεί να χρησιμοποιηθεί για να εξηγήσει αδιαφανή μοντέλα. Μέχρι την ημερομηνία συγγραφής αυτής της εργασίας, το πακέτο υποστηρίζει μοντέλα boosting, δέντρα αποφάσεων, αποφάσεις με λίστα κανόνα, γραμμικολογική παλινδρόμηση, εξήγηση SHAP kernel, SHAP δέντρο, LIME, morris ανάλυση ευαισθησίας, και μερική εξάρτηση. Διατίθεται στη διεύθυνση <https://github.com/interpretml/interpret>.
- Πακέτο IML [34], που συντηρείται από τον δημιουργό του Christoph Molnar. Το πακέτο καλύπτει ένα ευρύ φάσμα από μεθοδολογίες όπως χαρακτηριστικά σημασίας (feature importance), μερική εξάρτηση σχεδίασης, ατομική σχεδίαση προσδοκίας υπό όρους, συσσωρευμένα τοπικά εφέ, υποκατάστατα δέντρων, LIME, και SHAP. Διατίθεται στη διεύθυνση <https://github.com/christophM/iml>.

- Το πακέτο DeepExplain, που συντηρείται από τον Marco Ancona. Το πακέτο υποστηρίζει διάφορες τεχνικές gradient, όπως χάρτες ευκρίνειας, εισαγωγή gradient, ολοκληρωμένα gradient, DeepLIFT, LRP, κ.λπ. και όπως σύγκλιση, SHAP, κ.λπ. Διατίθεται στη διεύθυνση <https://github.com/marcoancona/DeepExplain>.
- Το DrWhy από την ModelOriented όπου είναι ένα πακέτο με διάφορες αγνωστικιστικές και εξειδικευμένες τεχνικές ETN για μοντέλα, όπως η σπουδαιότητα χαρακτηριστικών, τα ceteris paribus, η μερικής εξάρτησης σχεδίασης, η εξάρτηση υπό όρους, κλπ. Διατίθεται στη διεύθυνση <https://github.com/ModelOriented/DrWhy>.

5. Επίλογος και συμπεράσματα

Η τυφλή εμπιστοσύνη στα αποτελέσματα μιας πρόγνωσης ενός μοντέλου DL ή ML είναι αφελής, σύμφωνα πάντα με το σημερινά δεδομένα. Αυτό συμβαίνει λόγω της ισχυρής επίδρασης της μεροληψίας των δεδομένων, της αξιοπιστίας και των αποτελεσμάτων των αντιμαχόμενων παραδειγμάτων (adversarial networks) στην μηχανική μάθηση. Σε αυτήν την διπλωματική, ερευνήσαμε γιατί η ETN είναι σημαντική και ποιες είναι οι έννοιες που την περικλείουν, και τις κατηγοριοποιήσαμε σε σχέση με το πεδίο εφαρμογής, τη μεθοδολογία, τη χρήση και τη φύση τους. Δόθηκε έμφαση σε μεθοδολογίες που εφαρμόστηκαν στον τομέα της έξυπνης βιομηχανίας, η αλλιώς της Βιομηχανίας 4.0. Η βιομηχανία 4.0 αντιπροσωπεύει ένα σύγχρονο πρότυπο αυτοματισμού και επικοινωνίας δεδομένων στην κατασκευή και τις σχετικές βιομηχανίες. Συνδέει τεχνολογίες, συμπεριλαμβανομένων των κυβερνο-φυσικών συστημάτων, των αναλύσεων μεγάλων δεδομένων, του IoT, της υπολογιστικής νέφους και πολλών άλλων, και αναπτύσσει έξυπνες, πρακτικές εφαρμογές.

Από αυτήν την άποψη, η TN μπορεί να θεωρηθεί από τα σημαντικότερα στοιχεία της επανάστασης της Βιομηχανίας 4.0. Ωστόσο, οι μέθοδοι TN εξακολουθούν να αντιμετωπίζουν πολλές προκλήσεις. Πράγματι, τα συστήματα, οι μέθοδοι και οι αλγόριθμοι που βασίζονται κυρίως στην τεχνητή νοημοσύνη, είναι πεινασμένοι για δύναμη και χρειάζονται μια συνεχώς αυξανόμενη ποσότητα πυρήνων και Graphic Processing Units (GPUs) για να λειτουργήσουν αποτελεσματικά. Μια άλλη σημαντική πρόκληση που πρέπει να ληφθεί υπόψη είναι η άγνωστη φύση των μεθόδων που βασίζονται στην τεχνητή νοημοσύνη: πώς τα μοντέλα ML και DL προβλέπουν την παραγωγή; Είναι επίσης προκλητικό και απαιτητικό για τους ερευνητές να αποκτήσουν την ακρίβεια σε ανθρώπινο επίπεδο τέτοιων συστημάτων. Επιπλέον, οι μεθοδολογίες τεχνητής νοημοσύνης συχνά χρειάζονται τα ακόλουθα:

- 1) βελτιστοποίηση υπερπαραμέτρων
- 2) φινίρισμα
- 3) τεράστια σύνολα δεδομένων
- 4) ισχυρή υπολογιστική ικανότητα
- 5) συνεχής κατάρτιση στα δεδομένα.

Ως εκ τούτου, όλα τα μοντέλα ML και DL βασίζονται σε μια βασική πτυχή που είναι η διαθεσιμότητα των δεδομένων και των πόρων. Τέτοια δεδομένα συνήθως συλλέγονται και παράγονται από εκατομμύρια πηγές, συσκευές και χρήστες. υπάρχουν επίσης πιθανότητες ότι τα προσωπικά ή ευαίσθητα δεδομένα μπορούν να επηρεαστούν λόγω κυβερνοεπιθέσεων. Η μεροληπτική φύση των δεδομένων επηρεάζει επίσης την απόδοση του συστήματος που βασίζεται στην τεχνητή νοημοσύνη σε εφαρμογές πραγματικού χρόνου. Έτσι, η εγγενής ποιότητα της τεχνητής νοημοσύνης απαιτεί την προφανή εξήγηση που μπορεί να δώσει η ETN σε ένα σύστημα που επιτρέπει την ευφυΐα σε ανθρώπινο επίπεδο.

Τα ευρήματά της εργασίας έδειξαν ότι η σημαντική έρευνα στον ETN εστιάζεται σε μοντέλα-αγνωστικιστικών post-hoc αλγορίθμων επεξηγησιμότητας λόγω της ευκολότερης ενσωμάτωσής τους και της ευρύτερης εμβέλειάς τους. Το σημερινό ερευνητικό τοπίο στην αξιολόγηση της ETN δείχνει ότι το πεδίο της βρίσκεται ακόμη σε εξέλιξη και ότι οι μέθοδοι ETN πρέπει να αναπτυχθούν και να επιλεγούν με προσοχή. Μελέτες χρηστών έχουν δείξει ότι οι χάρτες τυπικών εξηγήσεων από μόνοι τους δεν μπορούν να βοηθήσουν στη λήψη αποφάσεων. Η ανθρώπινη προκατάληψη στην ερμηνεία των οπτικών εξηγήσεων θα μπορούσε να εμποδίσει τη σωστή χρήση της ETN σε κρίσιμες εφαρμογές. Οι πρόσφατες εξελίξεις στις αξιολογήσεις με βάση τον άνθρωπο δείχνουν πολλά υποσχόμενες βελτιώσεις στο τοπίο της αξιολόγησης ETN.

Πιο συγκεκριμένα τα συμπεράσματα που προκύπτουν μέσα από αυτή την εργασία είναι:

1. Οι περισσότερες τεχνικές που βρέθηκαν στη βιβλιογραφία ανήκουν στην κατηγορία των τοπικών εξηγήσεων δεδομένων, δηλαδή μπορούν να εξηγήσουν γιατί ο αλγόριθμος έλαβε αυτή την απόφαση για μια συγκεκριμένη είσοδο.
2. Αν και υπάρχουν πολλαπλές μέθοδοι ETN, με την δουλειά των [1] να συνοψίζει 29 τεχνικές, δεν επαρκούν από μόνες τους για να παρέχουν κατανοητές εξηγήσεις στους τελικούς χρήστες. Για το λόγο αυτό είναι ενσωματωμένες σε πλαίσια και εφαρμογές που παρέχουν ένα βολικό περιβάλλον διεπαφής και πρόσθετα πλαίσια για την επίτευξη αυτού του στόχου κατανόησης. Μέσα από την εργασία βρέθηκαν 6 τέτοια λογισμικά.
3. Στην έξυπνη βιομηχανία έχουν παρουσιαστεί αρκετές δουλειές που χρησιμοποιούν ETN για πιο κατανοητά αποτελέσματα από τους χειριστές. Μέσα από την βιβλιογραφική ανασκόπηση βρήκαμε δουλειές πάνω στην πρόβλεψη συντήρησης, και στον ποιοτικό έλεγχο. Για δεδομένα εικόνες πολύ δημοφιλές είναι οι προσεγγίσεις που δημιουργούν heatmaps όπως το Grad-CAM, ενώ για δεδομένα μορφής πίνακα έχουν χρησιμοποιηθεί διάφορες προσεγγίσεις. Μέσα από αυτήν την εργασία παρουσιάστηκαν οχτώ εφαρμογές την ETN σε διαφορετικές διεργασίες της έξυπνης βιομηχανίας.
4. Αν και η επεξηγηματική Τεχνητή Νοημοσύνη είναι ιδιαίτερα καινοτόμα και έχει τραβήξει το ερευνητικό κοινό, υπάρχει ακόμα ερευνητικό κενό και περιθώρια ανάπτυξης στο τομέα των μετρικών αξιολόγησης της επεξήγησης. Στην βιβλιογραφία βρέθηκαν 12 μετρικά αξιολόγησης που εφαρμόζονται για να αξιολογήσουν διαφορετικές έννοιες.

Πηγές

- [1] G. Sofianidis, J. M. Rožanec, D. Mladenicić, and D. Kyriazis, "A Review of Explainable Artificial Intelligence in Manufacturing." 2021.
- [2] A. C. Scott, W. J. Clancey, R. Davis, and E. H. Shortliffe, "Explanation capabilities of production-based consultation systems," *American Journal of Computational Linguistics*, pp. 1–50, 1977.
- [3] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5031–5042, 2022.
- [4] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.
- [5] E. Oztemel and S. Gursev, "Literature review of Industry 4.0 and related technologies," *J. Intell. Manuf.*, vol. 31, no. 1, pp. 127–182, Jan. 2020, doi: 10.1007/s10845-018-1433-8.
- [6] I. Yaqoob *et al.*, "Big data: From beginning to future," *International Journal of Information Management*, vol. 36, no. 6, Part B, pp. 1231–1247, 2016, doi: <https://doi.org/10.1016/j.ijinfomgt.2016.07.009>.

- [7] J. Li, "Cyber security meets artificial intelligence: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 12, pp. 1462–1474, Dec. 2018, doi: 10.1631/FITEE.1800573.
- [8] Z.-H. Zhou, *Machine learning*. Springer nature, 2021.
- [9] T. Tziolas, K. Papageorgiou, T. Theodosiou, E. Papageorgiou, T. Mastos, and A. Papadopoulos, "Autoencoders for anomaly detection in an industrial multivariate time series dataset," *Engineering Proceedings*, vol. 18, no. 1, p. 23, 2022.
- [10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [11] T. Tziolas *et al.*, "Wafer Map Defect Pattern Recognition using Imbalanced Datasets," in *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 2022, pp. 1–8. doi: 10.1109/IISA56318.2022.9904402.
- [12] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.
- [13] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.
- [14] K. Sokol and P. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 56–67.
- [15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
- [16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.
- [17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Jul. 2019, doi: 10.1007/s11263-019-01228-7.
- [18] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions." 2017.
- [19] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: <https://doi.org/10.1016/j.inffus.2019.12.012>.
- [20] I. Lage *et al.*, "An Evaluation of the Human-Interpretability of Explanation." 2019.
- [21] K. Sokol, A. Hepburn, R. Santos-Rodríguez, and P. A. Flach, "bLIMEy: Surrogate Prediction Explanations Beyond LIME," *CoRR*, vol. abs/1910.13016, 2019, [Online]. Available: <http://arxiv.org/abs/1910.13016>
- [22] C. Henin and D. Le Métayer, "A Generic Framework for Black-box Explanations," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 3667–3676. doi: 10.1109/BigData50022.2020.9377863.
- [23] V. Arya *et al.*, "AI Explainability 360 Toolkit," in *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, 2021, pp. 376–379.
- [24] J. M. Rožanec and D. Mladenović, "Semantic XAI for contextualized demand forecasting explanations," *arXiv preprint arXiv:2104.00452*, 2021.
- [25] P. Zajec, J. M. Rožanec, I. Novalija, B. Fortuna, D. Mladenović, and K. Kenda, "Towards active learning based smart assistant for manufacturing," in *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems: IFIP WG 5.7 International Conference, APMS 2021, Nantes, France, September 5–9, 2021, Proceedings, Part IV*, Springer, 2021, pp. 295–302.
- [26] C. V. Goldman, M. Baltaxe, D. Chakraborty, and J. Arinez, "Explaining learning models in manufacturing processes," *Procedia Computer Science*, vol. 180, pp. 259–268, 2021.
- [27] M. Lee, J. Jeon, and H. Lee, "Explainable AI for domain experts: a post Hoc analysis of deep learning for defect classification of TFT–LCD panels," *Journal of Intelligent Manufacturing*, pp. 1–13, 2021.
- [28] S. Yoo and N. Kang, "Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization," *Expert Systems with Applications*, vol. 183, p. 115430, 2021.

- [29] J.-R. Rehse, N. Mehdiyev, and P. Fettke, "Towards Explainable Process Predictions for Industry 4.0 in the DFKI-Smart-Lego-Factory," *KI - Künstliche Intelligenz*, vol. 33, no. 2, pp. 181–187, Jun. 2019, doi: 10.1007/s13218-019-00586-1.
- [30] E. Noh and S. Hong, "Automatic Screening of Bolts with Anti-Loosening Coating Using Grad-CAM and Transfer Learning with Deep Convolutional Neural Networks," *Applied Sciences*, vol. 12, no. 4, 2022, doi: 10.3390/app12042029.
- [31] C. Schockaert, V. Macher, and A. Schmitz, "VAE-LIME: deep generative model based approach for local data-driven model interpretability applied to the ironmaking industry," *arXiv preprint arXiv:2007.10256*, 2020.
- [32] B. Hrnjica and S. Softic, "Explainable AI in Manufacturing: A Predictive Maintenance Case Study," in *Advances in Production Management Systems. Towards Smart and Digital Manufacturing*, B. Lalic, V. Majstorovic, U. Marjanovic, G. von Cieminski, and D. Romero, Eds., Cham: Springer International Publishing, 2020, pp. 66–73.
- [33] O. Serradilla, E. Zugasti, J. Ramirez de Okariz, J. Rodriguez, and U. Zurutuza, "Adaptable and Explainable Predictive Maintenance: Semi-Supervised Deep Learning for Anomaly Detection and Diagnosis in Press Machine Data," *Applied Sciences*, vol. 11, no. 16, 2021, doi: 10.3390/app11167376.
- [34] C. Molnar, G. Casalicchio, and B. Bischl, "iml: An R package for interpretable machine learning," *Journal of Open Source Software*, vol. 3, no. 26, p. 786, 2018.