Det Humanistiske og Samfundsvidenskabelige Fakultet

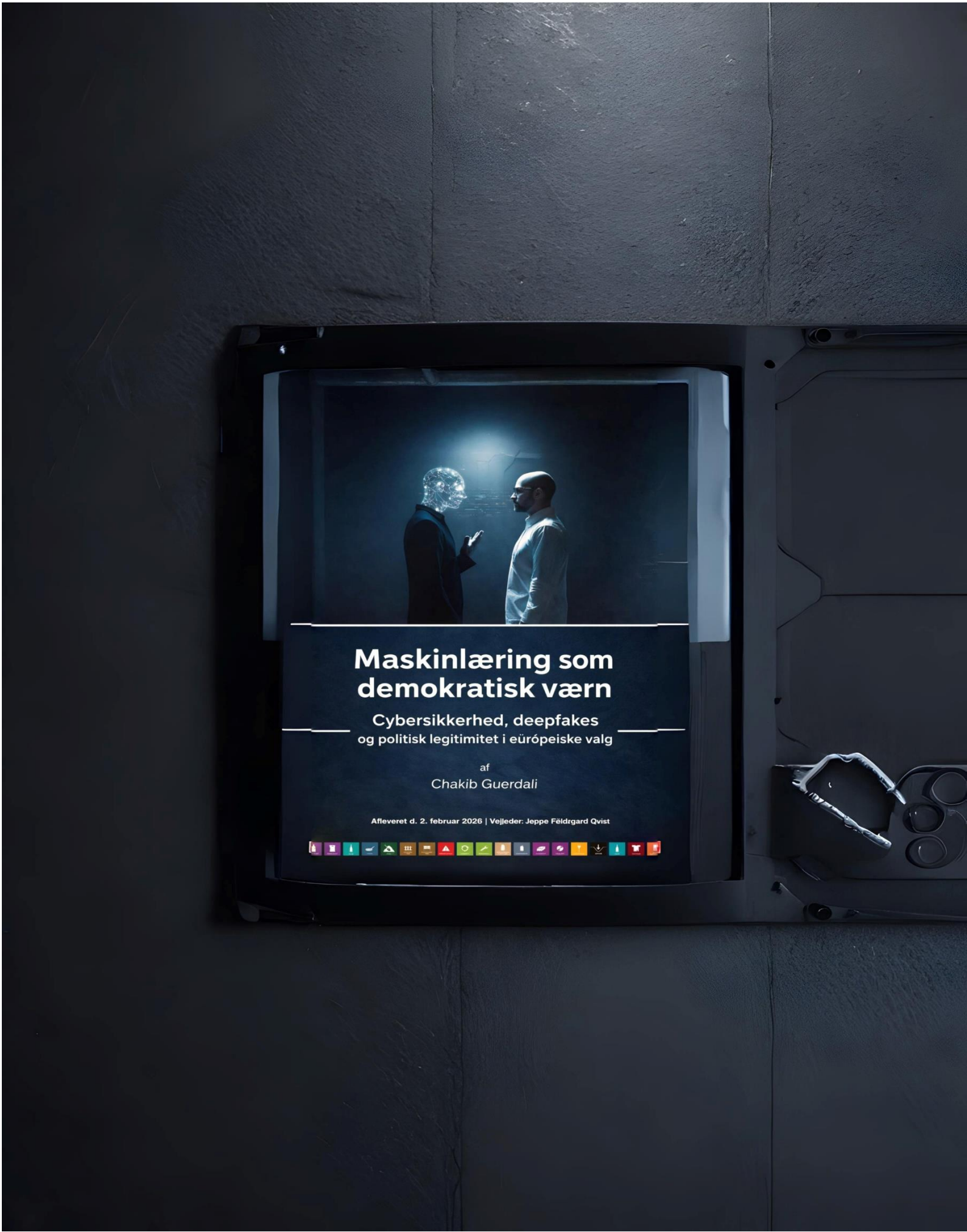
AALBORG UNIVERSITET

MASKINLÆRING SOM DEMOKRATISK VÆRN

CYBERSIKKERHED, DEEPFAKES OG POLITISK LEGITIMITET I EUROPÆISKE VALG

Speciale

Afleveringsfrist		Februar: 2026
Vejleder: Jeppe Fjeldgaard Qvist		**Institut:** AAU
Titel, dansk: Maskinlæring som demokratisk værn - Cybersikkerhed, deepfakes og politisk legitimitet i europæiske valg		
Titel, engelsk: Machine Learning as a Democratic Safeguard - Cybersecurity, Deepfakes, and Political Legitimacy in European Elections		
Besvarelsens antal typeenheder: 87.952 | 37 normal sider eller 52 sider		
Afleveret af:		
Fornavn: Chakib	**Efternavn:** Guerdali	**Studienummer:** 20204530


Maskinlæring som demokratisk værn
Cybersikkerhed, deepfakes og politisk legitimitet i eürópeiske valg
af
Chakib Guerdali
Afleveret d. 2. februar 2026 | Vejleder: Jeppe Fēldrgard Qvist

INDHOLD

ABSTRACT

Democratic elections constitute a foundational element of contemporary European societies and serve as the primary mechanism through which political legitimacy is established and reproduced. As political processes and public communication have become increasingly digitalised, electoral processes have simultaneously become more vulnerable to a wide range of digital threats. These threats include both traditional cyberattacks targeting election-related infrastructure and more complex forms of information manipulation, such as coordinated disinformation campaigns and the use of AI-generated media, commonly referred to as deepfakes. This development challenges not only the technical security of elections, but also raises fundamental questions concerning political legitimacy, citizens' trust, and the democratic space in Europe.

This thesis examines how machine learning is applied in practice by European election authorities and related actors to protect electoral processes from digital threats, and how the use of such technologies affects democratic legitimacy and public trust. The analysis adopts an interdisciplinary approach that integrates political theory, public administration, qualitative case study analysis, and technical insights from machine learning research. Empirically, the thesis is based on a comparative analysis of two European cases: the use of AI-generated deepfakes and disinformation during the Slovak parliamentary election in 2023, and cyberattacks targeting digital election-related infrastructure during the 2024 European Parliament elections. The two cases represent distinct but analytically comparable threat domains—information-based and infrastructural—and allow for an examination of how machine learning is deployed under different institutional and governance conditions.

The analysis demonstrates that machine learning already plays a significant role in safeguarding electoral processes, but that its application is uneven and highly context-dependent. In the infrastructural domain, machine learning-based cybersecurity solutions appear relatively mature and institutionally embedded, contributing to high output legitimacy by ensuring the stability and continuity of election administration under cyberattacks. In contrast, the use of machine learning to counter deepfakes and election-related disinformation remains far less institutionalised, resulting in fragmented governance structures and challenges to both input and throughput legitimacy.

The thesis further includes a technical demonstration involving the construction of a convolutional neural network for deepfake detection. Rather than aiming to optimise model performance, this demonstration is used reflexively to illustrate how machine learning models operate, as well as their limitations, including issues of bias, generalisation, probabilistic decision-making, and limited explainability. These technical characteristics have direct democratic implications, as they affect transparency, accountability, and the ability of citizens and decision-makers to understand and contest algorithmically supported decisions.

Overall, the thesis concludes that while machine learning can contribute substantially to protecting elections against digital threats, it does not constitute a neutral or sufficient solution in itself. Legitimate and effective use of machine learning in electoral contexts requires clear institutional frameworks, transparency, and meaningful human oversight. By analysing machine learning as a form of governance technology rather than merely a technical tool, this thesis contributes to existing research on AI, cybersecurity, and democracy, and highlights the normative trade-offs involved in deploying AI to safeguard democratic processes in Europe.

1. INTRODUKTION

Demokratiske valg udgør et fundamentalt element i moderne samfund og fungerer som den primære mekanisme, hvorigennem politisk legitimitet etableres og reproduceres. I takt med den stigende digitalisering af både samfundsliv og politiske processer er valgprocesser imidlertid blevet mere sårbare over for digitale trusler. Disse trusler spænder fra klassiske cyberangreb rettet mod valgrelateret infrastruktur til mere komplekse former for informationsmanipulation, herunder koordineret misinformation og brugen af kunstigt genererede medier, såkaldte deepfakes. Udviklingen udfordrer ikke blot den tekniske sikkerhed omkring valg, men rejser også grundlæggende spørgsmål om borgernes tillid, politisk legitimitet og det demokratiske råderum i bl.a. Europa.

Særligt kunstig intelligens eller AI og machine learning som i daglig tale kaldes ML, spiller en dobbeltrolle i denne udvikling. På den ene side muliggør fremskridt inden for generativ AI produktionen af overbevisende falsk indhold i et omfang og med en hastighed, der tidligere var utænkelig. Deepfake-videoer og -lyd kan anvendes til at udgive sig for politiske aktører, manipulere vælgernes opfattelser og skabe tvivl om, hvad der kan betragtes som autentisk information. På den anden side udgør AI og ML samtidig nogle af de mest centrale teknologiske redskaber til at opdage, analysere og modvirke netop disse trusler. Maskinlæringsbaserede systemer anvendes i stigende grad til at identificere anomalier i netværkstrafik, opdage koordinerede informationskampagner og analysere digitale medier for tegn på manipulation.

Denne dobbelte anvendelse af AI skaber et spændingsfelt, hvor teknologiske løsninger både fremstår som nødvendige for at beskytte demokratiske processer og som potentielle risikofaktorer i sig selv. Europæiske institutioner og nationale valgmyndigheder har i de senere år i stigende grad anerkendt denne problematik. Blandt andet har ENISA gentagne gange advaret om, at digitale trusler mod valgprocesser udgør en væsentlig sikkerhedsudfordring for EU's medlemsstater, særligt i lyset af AI-drevne informationsoperationer. Samtidig har EU vedtaget nye regulatoriske tiltag, herunder AI Act, som søger at etablere rammer for ansvarlig anvendelse af AI, herunder krav om transparens ved brug af syntetisk medieindhold i politiske sammenhænge.

På trods af disse initiativer eksisterer der fortsat betydelige videnshuller, særligt i relation til hvordan maskinlæring konkret anvendes af europæiske valgmyndigheder og tilknyttede aktører til

at beskytte valgprocesser i praksis. Meget af den eksisterende litteratur fokuserer enten på tekniske aspekter af AI-modeller eller på normative diskussioner om demokrati og digitalisering, men i mindre grad på samspillet mellem teknologi, forvaltning og demokratisk legitimitet. Der er derfor behov for analyser, der integrerer teknisk forståelse af AI og ML med politologiske og forvaltningsmæssige perspektiver.

Dette speciale placerer sig i dette krydsfelt mellem teknologi og demokrati. Formålet er at undersøge, hvordan maskinlæring anvendes som et cybersikkerhedsværktøj i europæiske valg, og hvilke konsekvenser denne anvendelse har for borgernes tillid og den demokratiske legitimitet. Særligt fokuseres der på to centrale typer af digitale trusler:

- Informationsbaserede trusler i form af deepfakes og misinformation og Infrastrukturelle trusler i form af cyberangreb mod valgrelaterede systemer og digitale tjenester.

Specialet anvender en kvalitativ casestudietilgang, hvor to europæiske cases analyseres for at belyse forskellige anvendelser af machine learning i beskyttelsen af valgprocesser. Casene er udvalgt, så de repræsenterer forskellige trusselsformer og organisatoriske kontekster, men samtidig muliggør en komparativ analyse. Analysen kombinerer dokumentstudier af offentlige rapporter, forskningslitteratur og policy-dokumenter med en teknisk demonstration af centrale ML-principper. Herunder indgår opbygningen af et konvolutionelt neuralt netværk eller CNN, der illustrerer, hvordan deepfake-detektion kan implementeres i praksis.

Ved at integrere tekniske AI-modeller i analysen søger specialet ikke at bidrage med ny maskinlæringsforskning, men derimod at skabe en dybere forståelse af, hvordan sådanne modeller fungerer, hvilke begrænsninger de har, og hvilke styringsmæssige og demokratiske implikationer deres anvendelse medfører. Dette er særligt relevant i en forvaltningskontekst, hvor beslutningstagere i stigende grad må forholde sig til teknologier, der opererer som såkaldte "black boxes", og hvor transparens, ansvarlighed og retssikkerhed udgør centrale normative hensyn.

Samlet set bidrager specialet til den voksende litteratur om digital demokratisikring ved at koble cybersikkerhed, AI og demokratisk legitimitet i en europæisk kontekst. Analysen sigter mod at vise, ML's vigtige rolle i beskyttelsen af valgprocesser, men at teknologien samtidig introducerer nye former for afhængighed, usikkerhed og styringsudfordringer. Disse spændinger gør det

nødvendigt at betragte AI som andet end blot et teknisk værktøj, men som et politisk og administrativt fænomen, der aktivt former rammerne for det demokratiske råderum i Europa.

2. PROBLEMFORMULERING

Anvendelsen af kunstig intelligens og machine learning i beskyttelsen af demokratiske valg indebærer et markant skift i, hvordan valgsikkerhed organiseres og udøves i Europa. Hvor valgprocesser traditionelt har været reguleret gennem juridiske rammer, organisatoriske procedurer og menneskelig kontrol, varetages centrale funktioner i dag i stigende grad af digitale systemer, der overvåger, klassificerer og reagerer på potentielle trusler i realtid.

Machine learning anvendes i denne sammenhæng til at identificere og håndtere et bredt spektrum af digitale trusler, herunder cyberangreb mod valgrelateret infrastruktur samt informationsbaserede trusler som misinformation og deepfakes. Disse teknologier muliggør en grad af automatisering og skalerbarhed, som er vanskelig at opnå gennem menneskelig overvågning alene, særligt i valgperioder præget af høj kompleksitet og tidskritiske beslutninger. Samtidig betyder anvendelsen af maskinlæringsbaserede systemer, at afgørende vurderinger i stigende grad baseres på statistiske sandsynligheder, mønstergenkendelse og beslutningstærskler fremfor eksplicitte juridiske kriterier.

Denne udvikling rejser et centralt politisk-administrativt problem. Når machine learning anvendes til at klassificere digital adfærd, indhold og trafik som legitim eller illegitim, flyttes dele af den normative vurdering af valgprocessers integritet fra mennesker til algoritmiske systemer. Dermed får tekniske designvalg – såsom valg af modeltype, træningsdata og klassifikationstærskler – direkte betydning for demokratiske værdier som proportionalitet, gennemsigtighed og politisk lighed. Samtidig kan manglende forklarbarhed og begrænset offentlig indsigt i disse systemers funktion udfordre borgernes tillid til både valgprocessen og de institutioner, der forvalter den. På trods af stigende politisk og institutionel opmærksomhed på digitale trusler mod valg – blandt andet fra europæiske sikkerheds- og reguleringsaktører – eksisterer der fortsat begrænset empirisk viden om, hvordan machine learning konkret anvendes i praksis i en valgkontekst, og hvordan denne anvendelse påvirker demokratisk legitimitet og borgernes tillid. Meget af den eksisterende litteratur

behandler enten de tekniske aspekter af maskinlæring eller de normative konsekvenser for demokratiet, men i mindre grad samspillet mellem teknologi, forvaltning og styring.

Dette speciale tager udgangspunkt i dette spændingsfelt og undersøger machine learning som både et teknisk sikkerhedsværktøj og en styringsteknologi i europæiske valgprocesser. Fokus er ikke på at optimere eller rangordne specifikke modeller, men på at analysere, hvordan maskinlæringsbaserede systemer anvendes, organiseres og legitimeres – og hvilke demokratiske konsekvenser denne anvendelse medfører.

På denne baggrund formuleres følgende overordnede problemformulering:

Hvordan anvendes maskinlæring konkret af europæiske valgmyndigheder til at beskytte valgprocesser mod digitale trusler som misinformation, cyberangreb og deepfakes – og hvordan påvirker denne anvendelse politisk legitimitet, borgernes tillid og det demokratiske råderum i Europa?

2.1 Afgrænsning og delspørgsmål

For at besvare problemformuleringen på en analytisk og struktureret måde operationaliseres den gennem følgende delspørgsmål:

- Hvordan anvendes machine learning teknologisk til at opdage og modvirke digitale trusler mod valgprocesser, herunder deepfakes, misinformation og cyberangreb?
  Dette delspørgsmål fokuserer på de tekniske og funktionelle aspekter af machine learning, herunder forskellen mellem supervised og unsupervised learning samt konkrete anvendelser i en valgkontekst.
- Hvordan er anvendelsen af machine learning organisatorisk og institutionelt forankret hos europæiske valgmyndigheder og relaterede aktører?
  Her analyseres samspillet mellem offentlige myndigheder, private teknologileverandører og digitale platforme samt de styringsmæssige udfordringer, der opstår i denne konstellation.

- Hvilke konsekvenser har brugen af machine learning i valgsikkerhed for politisk legitimitet, borgernes tillid og det demokratiske råderum?
  Dette delspørgsmål adresserer de normative og demokratiske implikationer af algoritmisk understøttet valgsikkerhed, herunder spørgsmål om transparens, proportionalitet, fairness og demokratisk kontrol.

Ved at kombinere disse delspørgsmål søger specialet at udvikle en helhedsforståelse af machine learning som både et teknisk sikkerhedsværktøj og et politisk-administrativt fænomen, der aktivt former rammerne for demokratisk deltagelse og legitimitet i Europa.

3. TEORETISK RAMME

Dette speciale undersøger anvendelsen af machine learning i beskyttelsen af europæiske valgprocesser og de demokratiske implikationer heraf. For at kunne analysere både de teknologiske og politisk-administrative dimensioner af problemstillingen er det nødvendigt at anvende en tværgående teoretisk ramme, der kombinerer perspektiver fra demokratiteori, offentlig forvaltning og studier af digital og algoritmisk governance. Den teoretiske ramme har således til formål at etablere de begrebslige redskaber, der gør det muligt at analysere, hvordan maskinlæring både fungerer som et sikkerhedsværktøj og som et styringsinstrument, der påvirker politisk legitimitet, borgernes tillid og det demokratiske råderum (Yeung, 2018, s. 507–509; Broeders, D., & Sukumar, A., 2024.

Afsnittet er opbygget i fire dele. Først præsenteres et legitimationsperspektiv med fokus på input-, throughput- og output-legitimitet. Dernæst diskuteres borgernes tillid i relation til institutioner og teknologier. Herefter introduceres begrebet algoritmisk governance som en ramme for at forstå delegation af beslutningskompetence til teknologiske systemer. Afslutningsvis anvendes securitization theory til at analysere, hvordan digitale trusler mod valg konstrueres som sikkerhedsproblemer, og hvilke konsekvenser dette har for demokratiske processer.

3.1 Demokratisk legitimitet i en digital kontekst

Politisk legitimitet udgør et centralt analytisk begreb i specialet, idet valgprocesser netop fungerer som den primære mekanisme, hvorigennem legitimitet etableres i demokratiske systemer. I dette speciale anvendes en flerdimensionel forståelse af legitimitet, der tager udgangspunkt i en opdeling mellem input-, throughput- og output-legitimitet (Scharpf, 1999, s. 6–13; Schmidt, 2013, s. 2–5).

Input-legitimitet refererer til borgernes mulighed for at deltage i politiske processer på lige og frie vilkår. I en valgkontekst handler dette om adgang til korrekt information, fravær af manipulation og mulighed for at danne politiske præferencer uden skjult påvirkning. Deepfakes og koordineret misinformation udgør her en direkte trussel, idet de kan forvride informationsgrundlaget og dermed underminere den demokratiske deltagelse (Bennett & Livingston, 2018, s. 124–127; Chesney & Citron, 2019, s. 147–150). Anvendelsen af machine learning til at opdage og modvirke sådanne

trusler kan i denne forstand bidrage positivt til input-legitimiteten, hvis teknologien effektivt reducerer manipulation.

Throughput-legitimitet omhandler kvaliteten af de processer, hvorigennem politiske beslutninger træffes og implementeres. Centralt her står principper som transparens, ansvarlighed og retssikkerhed (Scharpf, 1999, s. 10–11). Når valgmyndigheder anvender machine learning-baserede systemer til overvågning, filtrering eller beslutningsstøtte, kan throughput-legitimiteten udfordres, særligt hvis disse systemer fungerer som såkaldte "black boxes". Manglende indsigt i, hvordan algoritmiske klassifikationer foretages, kan gøre det vanskeligt for både borgere og beslutningstagere at efterprøve og anfægte afgørelser, hvilket potentielt svækker den demokratiske kontrol (Burrell, 2016, s. 3–6; Yeung, 2018, s. 510–511).

Output-legitimitet relaterer sig til systemets evne til at levere effektive og ønskede resultater. I denne sammenhæng handler det om, hvorvidt machine learning faktisk formår at beskytte valgprocesser mod digitale trusler. En høj grad af teknisk effektivitet kan styrke output-legitimiteten ved at sikre stabile og troværdige valg, men denne legitimitet er skrøbelig, hvis den ikke ledsages af transparens og respekt for demokratiske rettigheder (Schmidt, 2013, s. 7–9). Et centralt analytisk spørgsmål i specialet er derfor, om styrket output-legitimitet gennem teknologisk effektivitet sker på bekostning af input- og throughput-legitimitet.

3.2 Borgernes tillid og teknologisk mediation

Tillid er tæt forbundet med legitimitet, men udgør et selvstændigt analytisk begreb. Borgernes tillid til valgprocesser afhænger ikke alene af formelle regler og institutionelle garantier, men også af oplevelsen af fairness, forståelighed og kontrol (Levi & Stoker, 2000, s. 476–478). I takt med at teknologier spiller en større rolle i valgadministrationen, bliver tillid i stigende grad medieret gennem tekniske systemer (Vaccari & Chadwick, 2020, s. 10–13).

I dette speciale skelnes der mellem institutionel tillid og teknologisk tillid. Institutionel tillid refererer til borgernes tillid til valgmyndigheder, domstole og andre demokratiske institutioner. Teknologisk tillid refererer derimod til borgernes tillid til, at de teknologiske systemer, som institutionerne anvender, fungerer korrekt, retfærdigt og uden skjulte dagsordener (Kitchin, 2017, s. 17–19). Disse

to former for tillid er tæt forbundne, idet svigt i teknologiske systemer hurtigt kan oversættes til mistillid til institutionerne bag dem.

Anvendelsen af machine learning i valgsikkerhed kan på den ene side styrke tilliden ved at forhindre synlige angreb og manipulation. På den anden side kan usynlige og komplekse teknologiske processer skabe usikkerhed og mistro, særligt hvis borgere oplever, at beslutninger træffes på baggrund af systemer, de ikke forstår. Denne problematik er særlig relevant i forbindelse med automatiseret indholdsmoderation, hvor fejlagtige klassifikationer kan opfattes som politisk censur, selv når de er teknisk begrundede (Chesney & Citron, 2019, s. 155–158).

Tillid bliver dermed ikke blot et spørgsmål om teknisk korrekthed, men også om kommunikation, forklarbarhed og institutionel ansvarlighed (Rudin, 2019, s. 206–210). Dette speciale anvender tillidsbegrebet til at analysere, hvordan machine learning påvirker borgernes oplevelse af valgprocessens troværdighed, både direkte gennem teknologisk funktionalitet og indirekte gennem institutionel praksis.

3.3 Algoritmisk governance og delegation af beslutningskompetence

For at forstå machine learning som et styringsredskab anvender specialet begrebet algoritmisk governance. Dette begreb refererer til styringsformer, hvor algoritmiske systemer anvendes til at træffe, understøtte eller strukturere beslutninger, som tidligere blev håndteret af mennesker (Yeung, 2018, s. 507–508; Katzenbach & Ulbricht, 2019, s. 2–4).

Algoritmisk governance indebærer en form for delegation af beslutningskompetence, hvor mennesker overlader dele af vurderings- og beslutningsprocessen til teknologiske systemer. Denne delegation er ofte begrundet i effektivitet, skalerbarhed og hastighed, men rejser samtidig spørgsmål om ansvar og kontrol (Kitchin, 2017, s. 18–20). Hvis en machine learning-model fejlagtigt klassificerer legitim politisk kommunikation som misinformation, hvem bærer da ansvaret? Valgmyndigheden, softwareleverandøren eller modellen selv?

Et centralt aspekt ved algoritmisk governance er spændingen mellem automatisering og menneskelig dømmekraft. I praksis anvendes machine learning ofte som beslutningsstøtte snarere end som fuldt automatiserede beslutningstagere. Ikke desto mindre kan systemernes anbefalinger få betydelig indflydelse, særligt i tidskritiske situationer som valg (Selbst et al., 2019, s. 61–64).

Specialet anvender begrebet algoritmisk governance til at analysere, hvordan machine learning integreres i organisatoriske beslutningsprocesser, og hvordan dette påvirker demokratiske principper om ansvarlighed og gennemsigtighed.

3.4 Securitization af valg og digitale trusler

Endelig inddrager specialet securitization theory for at analysere, hvordan digitale trusler mod valg italesættes og håndteres politisk. Ifølge securitization theory bliver et fænomen et sikkerhedsproblem, når det fremstilles som en eksistentiel trussel, der legitimerer ekstraordinære foranstaltninger (Buzan, Wæver & de Wilde, 1998, s. 23–26).

I de senere år er valgprocesser i stigende grad blevet securitiseret, særligt i relation til cyberangreb, udenlandsk indblanding og informationsoperationer (Broeders, D., & Sukumar, A., 2024). Denne securitisering har haft betydning for legitimeringen af teknologiske indgreb. Når valg fremstilles som truede af alvorlige og akutte digitale risici, kan det retfærdiggøre anvendelsen af omfattende overvågnings- og filtreringssystemer baseret på machine learning (Saeva & Tasheva, 2024, s. 221–223).

Samtidig kan securitiseringen føre til en normalisering af midlertidige undtagelsesforanstaltninger, der på længere sigt kan påvirke det demokratiske råderum (Buzan et al., 1998, s. 29–30). Securitization theory anvendes i specialet til at analysere, hvordan diskurser om cybersikkerhed og AI skaber politisk opbakning til teknologiske løsninger, og hvordan dette påvirker balancen mellem sikkerhed og demokratiske rettigheder.

Den teoretiske ramme kombinerer legitimitetsteori, tillid, algoritmisk governance og securitization theory for at skabe et helhedsorienteret analytisk grundlag. Samlet set muliggør denne ramme en analyse af machine learning som et effektivt værktøj i cybersikkerhed og et styringsmæssigt fænomen med betydelige demokratiske implikationer (Yeung, 2018, s. 511; Schmidt, 2013, s. 10–11).

I de følgende kapitler anvendes denne teoretiske ramme systematisk til at analysere de udvalgte cases og vurdere, hvordan brugen af machine learning påvirker valgprocesser, politisk legitimitet og borgernes tillid i Europa.

4. METODOLOGI

Dette speciale anvender en kvalitativ forskningsstrategi med henblik på at analysere, hvordan machine learning anvendes til at beskytte europæiske valgprocesser mod digitale trusler, samt hvilke demokratiske implikationer denne anvendelse har. Metodevalget er styret af problemformuleringens karakter, som ikke alene fokuserer på teknologisk funktionalitet, men i lige så høj grad på politiske, institutionelle og normative konsekvenser. Specialet kombinerer derfor en kvalitativ casestudietilgang med en teknisk demonstration af centrale maskinlæringsprincipper for at opnå en helhedsorienteret forståelse af fænomenet (Yin, 2018, s. 15–18; Flyvbjerg, 2006, s. 219–223).

Afsnittet redegør først for det overordnede forskningsdesign og valg af kvalitativ metode. Dernæst præsenteres casestudiedesignet og begrundelsen for case udvælgelsen. Herefter redegøres der for datagrundlaget og analysestrategien, herunder anvendelsen af dokumentanalyse og teknisk modellering. Afslutningsvis diskuteres metodiske begrænsninger samt specialets validitet og reliabilitet.

4.1 Forskningsdesign og metodisk tilgang

Specialets forskningsdesign er eksplorativt og analytisk. Formålet er ikke at teste en kausal hypotese om effekten af machine learning på valgresultater, men derimod at analysere, hvordan teknologien anvendes i praksis, og hvordan denne anvendelse kan forstås i relation til demokratisk legitimitet, borgernes tillid og det demokratiske råderum. Dette taler for en kvalitativ tilgang, der muliggør dybdegående analyse af komplekse sammenhænge mellem teknologi, forvaltning og demokrati (Yin, 2018, s. 8–10).

Den kvalitative tilgang gør det muligt at undersøge machine learning som et socialt og institutionelt fænomen, fremfor udelukkende som en menneskeskabt genstand. Teknologien betragtes således ikke som neutral, men som indlejret i organisatoriske praksisser, politiske diskurser og normative

rammer (Selbst et al., 2019, s. 61–64). Dette metodiske udgangspunkt er centralt for at forstå, hvorfor samme teknologiske værktøj kan have forskellige demokratiske konsekvenser afhængigt af kontekst og anvendelse.

4.2 Casestudiedesign

Specialet anvender et kvalitativt casestudiedesign baseret på to europæiske cases. Casestudier er særligt velegnede til at analysere komplekse og kontekstafhængige fænomener, hvor grænserne mellem fænomen og kontekst er uklare (Yin, 2018, s. 16–17). I dette tilfælde er anvendelsen af machine learning tæt forbundet med både institutionelle strukturer, teknologiske infrastrukturer og politiske diskurser.

De to cases er udvalgt ud fra en strategisk case udvælgelse, hvor formålet er at belyse forskellige typer af digitale trusler og forskellige former for machine learning-anvendelse (Flyvbjerg, 2006, s. 230–232):

Case 1 fokuserer på anvendelsen af machine learning til at opdage og håndtere deepfakes og misinformation i forbindelse med nationale valg i Europa.

Case 2 fokuserer på anvendelsen af machine learning til at beskytte valgrelateret digital infrastruktur mod cyberangreb, herunder DDoS-angreb, i forbindelse med europæiske valg.

Casene repræsenterer således to forskellige, men analytisk sammenlignelige trusselsformer: informationsbaserede trusler og infrastrukturelle trusler. Denne opdeling muliggør en komparativ analyse af, hvordan machine learning anvendes på tværs af forskellige sikkerhedsdimensioner, og hvordan dette påvirker demokratiske værdier på forskellig vis (George & Bennett, 2005, s. 67–70).

4.3 Caseudvælgelse og afgrænsning

Case udvælgelsen er baseret på et princip om analytisk variation fremfor repræsentativitet (Flyvbjerg, 2006, s. 229–230). Formålet er ikke at generalisere statistisk til alle europæiske valg, men at identificere mekanismer og mønstre, som kan have bredere analytisk relevans.

Begge cases er udvalgt, fordi de opfylder følgende kriterier:

- Der har været dokumenterede digitale trusler mod valgprocessen.
- Machine learning-baserede teknologier har spillet en rolle i håndteringen af disse trusler.
- Der foreligger offentligt tilgængeligt materiale i form af rapporter, analyser og dokumentation (Bowen, 2009, s. 28–30).

Afgrænsningen indebærer, at specialet ikke analyserer alle former for AI-anvendelse i valg, f.eks. elektronisk stemmeafgivning eller automatiseret stemmeoptælling. Fokus er udelukkende på machine learning som et værktøj til cybersikkerhed og informationsbeskyttelse.

4.4 Datagrundlag og dokumentanalyse

Specialets empiriske grundlag består primært af dokumentdata. Dette inkluderer:

- Offentlige rapporter fra europæiske institutioner og nationale myndigheder, herunder analyser fra ENISA.
- Policy-dokumenter og strategipapirer om valgsikkerhed og AI.
- Akademiske forskningsartikler om deepfakes, misinformation og cyberangreb.
- Dokumentation og tekniske beskrivelser af machine learning-modeller og datasæt.

Dokumentanalysen anvendes til at identificere, hvordan digitale trusler italesættes, hvilke teknologiske løsninger der anvendes, og hvilke normative hensyn der fremhæves (Bowen, 2009, s. 32–35). Analysen fokuserer både på indhold og kontekst, herunder hvilke aktører der producerer dokumenterne, og hvilke interesser og antagelser der ligger bag.

4.5 Teknisk demonstration som analytisk metode

Ud over dokumentanalysen inddrager specialet en teknisk demonstration i form af opbygningen af et konvolutionelt neuralt netværk (CNN) til deepfake-detektion. Denne demonstration fungerer ikke som et eksperiment med henblik på at opnå optimal performance, men som et analytisk

redskab til at illustrere, hvordan machine learning-modeller fungerer i praksis (Burrell, 2016, s. 3–5).

Den tekniske del har tre metodiske formål:

- At konkretisere abstrakte beskrivelser af machine learning.
- At synliggøre modellernes begrænsninger, herunder risikoen for bias, falske positive og manglende forklarbarhed.
- At skabe grundlag for en informeret diskussion af governance- og legitimitetsimplikationer (Rudin, 2019, s. 206–210).

Ved at kombinere teknisk modellering med kvalitativ analyse bidrager specialet til at bygge bro mellem teknologisk og samfundsvidenskabelig viden (Selbst et al., 2019, s. 61–64).

4.6 Analysestrategi

Analysen gennemføres i tre trin. Først analyseres hver case individuelt med fokus på trusselsbilledet, den anvendte teknologiske løsning og den institutionelle kontekst. Dernæst sammenlignes casene med henblik på at identificere ligheder og forskelle i anvendelsen af machine learning på tværs af organisatoriske og politiske rammer. Endelig diskuteres resultaterne i lyset af den teoretiske ramme med særligt fokus på demokratisk legitimitet, borgernes tillid og algoritmisk governance (George & Bennett, 2005, s. 75–78).

Denne analytiske tilgang muliggør en systematisk kobling mellem empiriske observationer og teoretiske begreber og understøtter dermed en nuanceret besvarelse af specialets problemformulering.

4.7 Metodisk refleksion og begrænsninger

Dette speciale anvender en kvalitativ forskningsstrategi kombineret med en teknisk demonstration af maskinlæring. Det metodiske valg er truffet på baggrund af problemformuleringens karakter, idet specialet ikke søger at måle effekten af machine learning kvantitativt, men derimod at analysere,

hvordan teknologien anvendes i praksis, og hvilke demokratiske og styringsmæssige implikationer denne anvendelse medfører (Flyvbjerg, 2006, s. 221–223).

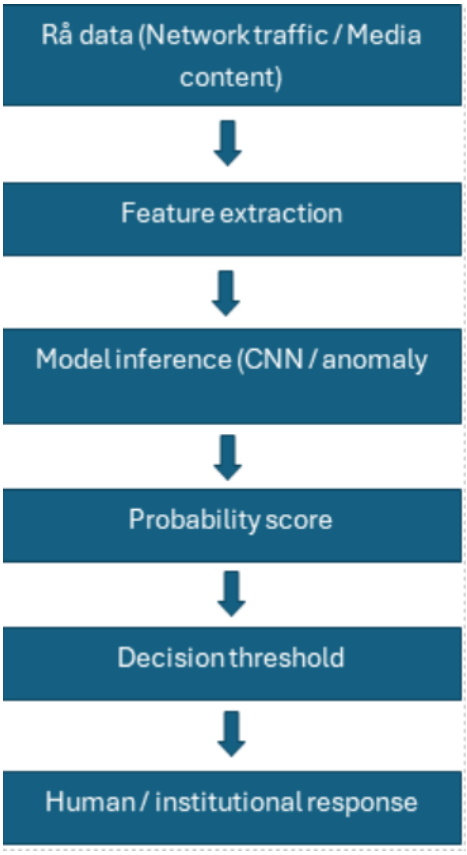
En kvalitativ tilgang er særligt velegnet i undersøgelsen af machine learning i valgkontekster, da teknologien ikke alene fungerer som et teknisk system, men som et politisk-administrativt redskab, der indgår i komplekse institutionelle og normative sammenhænge (Selbst et al., 2019, s. 62–64). Anvendelsen af casestudier muliggør en dybdegående analyse af konkrete anvendelser af machine learning og er velegnet til studiet af nye og komplekse fænomener, hvor eksisterende teori ikke tilbyder entydige forklaringer (Yin, 2018, s. 4–6).

Den tekniske demonstration indgår ikke som et empirisk performance-studie, men som et analytisk redskab, der bidrager til begrebsmæssig validitet og understøtter en informeret analyse af forklarbarhed, ansvarlighed og risikoen for fejlagtige klassifikationer (Burrell, 2016, s. 6–8; Rudin, 2019, s. 206–210). Demonstrationen er bevidst forenklet og kan ikke repræsentere den fulde kompleksitet i operative deepfake- eller cybersikkerhedssystemer, men fungerer som et illustrativt middel til at synliggøre centrale mekanismer og afvejninger.

Specialet er samtidig underlagt en række metodiske begrænsninger. Analysen baserer sig primært på sekundære kilder, hvilket indebærer en afhængighed af eksisterende dokumentation og rapportering. Derudover muliggør den kvalitative tilgang ikke fastlæggelse af kausale effekter mellem anvendelsen af machine learning og konkrete valgudfald (Yin, 2018, s. 42–45). Fravalg af interviews og kvantitative eksperimenter er et bevidst valg, der afspejler specialets analytiske fokus og disciplinære forankring inden for politik og administration (Bowen, 2009, s. 33–35).

Specialets validitet er således primært analytisk og teoretisk, mens generaliserbarheden er analytisk snarere end statistisk (Flyvbjerg, 2006, s. 224–226). Potentielle bias relateret til anvendelsen af sekundære kilder og forenklet teknisk modellering adresseres gennem metodisk transparens og eksplicit refleksion, hvilket styrker efterprøvelighed og analytisk stringens.

5. MACHINE LEARNING OG KUNSTIG INTELLIGENS SOM VÆRKTØJ TIL VALGSIKKERHED



Figur 5.1: Overordnet pipeline for anvendelse af machine learning i valgsikkerhed. Figuren illustrerer, hvordan rå data behandles gennem feature-ekstraktion og modelinferens, hvorefter probabilistiske output omsættes til beslutninger via organisatoriske

Dette kapitel har til formål at give en teknisk funderet, men analytisk orienteret redegørelse for, hvordan machine learning kan anvendes til at opdage og modvirke digitale trusler mod valgprocesser. Kapitlet fungerer som bindeled mellem specialets teoretiske ramme og de efterfølgende empiriske analyser ved at konkretisere, hvordan abstrakte begreber som algoritmisk governance, automatisering og teknologisk delegation materialiserer sig i faktiske maskinlæringsmodeller (Kitchin, 2017, s. 17–20; Yeung, 2018, s. 507–510).

Fokus er ikke på at optimere modelperformance, men på at demonstrere de grundlæggende principper bag machine learning-systemer, der anvendes i praksis til deepfake-detektion og cybersikkerhed. Kapitlet indledes med en gennemgang af forskellen mellem supervised og unsupervised learning. Herefter præsenteres opbygningen af et konvolutionelt neuralt netværk (CNN) som en konkret teknisk demonstration. Afslutningsvis diskuteres modellens begrænsninger og de demokratiske implikationer af teknisk automatisering.

Afsnittet integrerer tre analytiske elementer:

- En visuel fremstilling af modellens arkitektur (Figur 5.1)
- En konceptuel ROC-kurve som beslutningsværktøj i maskinlæringsbaseret valgsikkerhed (Figur 5.2), og

- En sammenlignende analyse af modelperformance, forklarbarhed og fejlprofiler for henholdsvis en forklarlig baseline-model og en kompleks 1D-CNN-model (Tabel 5.2, Tabel 5.3 og Tabel 5.4).

5.1 Teknisk demonstration: Modellens arkitektur og funktion

Lagtype	*Parametre*	*Funktion*
Input	81 x 1	Sekventielle netværksdata
Conv1D	64 filtre, kernel k	Lokale mønstre
BatchNorm	-	Stabilisering
ReLU	-	Ikke-linearitet
MaxPooling	-	Tidsreduktion
Conv1D	128 filtre	Mellemlagsrepræsentation
Conv1D	256 filtre	Strukturelle mønstre
Global Avg Pool	-	Aggregering
Dense	128 neuroner	Beslutningslag
Output	Sigmoid	Sandsynlighed

Tabel 5.1: Overordnet arkitektur for den anvendte 1D-konvolutionelle neurale netværksmodel. Figuren viser informationsflowet fra sekventielt input gennem tre konvolutionelle blokke med stigende kompleksitet til global aggregering og binær klassifikation.

Som vist i Tabel 5.1 illustrerer den visuelle fremstilling den overordnede arkitektur for den 1D-konvolutionelle neurale netværksmodel, der anvendes i dette speciale som en teknisk demonstration. Figuren fungerer som en konceptuel oversigt over modellens opbygning og informationsflow og har ikke til formål at dokumentere modellens performance eller træningsresultater.

Formålet med den visuelle fremstilling er at give læseren en intuitiv forståelse af, hvordan sekventielle inputdata behandles gennem modellens forskellige lag, fra rå input til endelig klassifikationsbeslutning. Som illustreret i Tabel 5.1 viser modellen, hvordan mønstre i data operationaliseres gennem arkitektoniske valg fremfor eksplicit regelbaseret modellering.

Som det fremgår af Tabel 5.1, modtager modellen sekventielle inputdata med dimensionen 81 × 1, hvilket repræsenterer et tidsafhængigt observationsvindue. Inputdata behandles herefter gennem tre konvolutionelle blokke med stigende kompleksitet, som successivt udtrækker mere abstrakte repræsentationer af mønstre i data, samtidig med at den tidslige dimension reduceres gennem pooling-operationer.

Efter feature-ekstraktionen aggregeres informationen, som vist i Tabel 5.1, ved hjælp af global average pooling, hvor de tidslige repræsentationer sammenfattes til en samlet, fast-dimensionel repræsentation. Denne repræsentation anvendes efterfølgende i fuldt forbundne lag til at foretage den endelige binære klassifikation.

Det er centralt at understrege, at Tabel 5.1 ikke skal forstås som et performance-bevis, men som en metodisk visualisering af modellens logik og funktion. Figuren bidrager dermed til at gøre en ellers kompleks "black-box"-model mere analytisk tilgængelig, hvilket er særligt relevant i en social data science-kontekst, hvor refleksiv og ansvarlig anvendelse af teknologiske metoder er centrale hensyn.

5.1.1 NEURAL NETVÆRKSMODEL (CNN)

I dette studie anvendes et 1-dimensionelt konvolutionelt neuralt netværk (1D-CNN) til binær klassifikation af sekventielle netværksdata med henblik på at identificere distribuerede denial-of-service-angreb (DDoS). Modellens arkitektur er illustreret i Tabel 5.1og gennemgås nedenfor.

Modellens input består af sekvenser med længde 81 og én kanal, hvor hver observation repræsenterer et tidsafhængigt mønster af aggregerede netværksfeatures. Denne repræsentation gør det muligt at analysere dynamiske trafikforløb, hvor relationer mellem observationer over tid er centrale.

Den første del af modellen består af tre konvolutionelle blokke. Det første konvolutionelle lag anvender 64 filtre og har til formål at identificere lokale og lav-niveau mønstre i inputsekvensen, såsom gentagelser, pludselige stigninger eller regelmæssige variationer i trafikken. Efterfølgende batch-normalisering stabiliserer træningen ved at normalisere aktiveringerne, mens ReLU-aktivering introducerer ikke-linearitet og muliggør læring af komplekse sammenhænge.

Efter den første og anden konvolutionelle blok anvendes max-pooling, som reducerer den tidslige dimension. Dette bidrager til at reducere modellens kompleksitet og øger robustheden over for mindre tidsforskydninger i mønstrene. De efterfølgende konvolutionelle lag øger antallet af filtre til henholdsvis 128 og 256 og lærer mere sammensatte og strukturelle repræsentationer af netværkstrafikken.

Efter feature-ekstraktionen anvendes global average pooling, som sammenfatter hver feature-kanal ved at beregne gennemsnittet over hele sekvensen. Dette resulterer i en fast længde repræsentation og reducerer antallet af parametre sammenlignet med traditionelle flatten-lag, hvilket mindsker risikoen for overfitting.

Den aggregerede feature-vektor føres videre til et fuldt forbundet lag med 128 neuroner, som fungerer som et beslutningslag, hvor information fra de lærte features integreres. Dropout anvendes som regularisering for at forbedre modellens generaliseringsevne. Output-laget består af én neuron med sigmoid-aktivering, som producerer en sandsynlighed for, at den pågældende observation tilhører den positive klasse (DDoS).

Modellen anvendes i dette speciale som en illustrativ teknisk demonstration og ikke som et produktionsklart detektionssystem. Samlet følger modellen en hierarkisk læringsstruktur, hvor tidlige lag lærer simple lokale mønstre, mellemlag lærer sammensatte strukturer, og sene lag aggregerer informationen til en global repræsentation, der anvendes til klassifikation. Set i et social data science-perspektiv illustrerer modellen, hvordan kollektive og strukturelle adfærdsmønstre kan identificeres automatisk ud fra sekventielle data uden eksplicit regelbaseret modellering.

5.2 Supervised og unsupervised i en valgsikkerhedskontekst

Machine learning kan overordnet opdeles i supervised og unsupervised learning, som adskiller sig ved, hvordan modeller trænes, og hvilke problemtyper de er velegnede til at løse (Goodfellow, Bengio & Courville, 2016, s. 98–104).

Supervised learning baserer sig på træning med mærkede data, hvor hvert datapunkt er forbundet med en kendt label. I en valgsikkerhedskontekst kan dette eksempelvis være digitale medier, der på forhånd er klassificeret som enten autentiske eller manipulerede. Modellen lærer at identificere

statistiske mønstre i data, som korrelerer med de givne labels, og kan herefter anvendes til at klassificere nye, ukendte eksempler (Goodfellow et al., 2016, s. 102–106).

Supervised learning er særligt velegnet til deepfake-detektion, fordi problemstillingen netop kan formuleres som en binær klassifikationsopgave. Samtidig indebærer denne tilgang en afhængighed af kvaliteten og repræsentativiteten af træningsdata, hvilket gør modellen sårbar over for bias og overfitting (Dolhansky et al., 2020, s. 4–7).

Unsupervised learning adskiller sig ved, at modellen trænes på umærkede data uden kendte labels. I stedet søger modellen at identificere underliggende strukturer, mønstre eller afvigelser i data. Denne tilgang anvendes ofte i cybersikkerhed til anomali-detektion, hvor målet er at identificere usædvanlig adfærd i eksempelvis netværkstrafik, der kan indikere et cyberangreb (Sommer & Paxson, 2010, s. 22–25).

I valgkontekster anvendes unsupervised learning særligt til at opdage nye eller ukendte angrebstyper. Denne fleksibilitet sker dog på bekostning af præcision og forklarbarhed, idet det kan være vanskeligt at afgøre, hvorfor et givent datapunkt klassificeres som en anomali (ENISA, 2024, s. 18–21).

Denne skelnen mellem supervised og unsupervised learning danner grundlag for at forstå deepfake-detektion som et klassifikationsproblem, hvilket uddybes i det følgende afsnit.

5.3 Deepfake-detektion som klassifikationsproblem

Deepfakes udgør en særlig udfordring i valgkontekster, fordi de direkte angriber informationsgrundlaget for demokratisk deltagelse (Chesney & Citron, 2019, s. 147–150). Teknisk set kan deepfake-detektion forstås som et mønstergenkendelsesproblem, hvor målet er at identificere subtile statistiske afvigelser mellem autentisk og manipuleret medieindhold (Dolhansky et al., 2020, s. 3–6).

I en valgkontekst har fejlklassifikation særligt alvorlige implikationer, idet både falske positive og falske negative kan påvirke borgernes tillid til politisk information og legitimiteten af myndigheders indgreb.

Tidlige deepfake-modeller efterlod ofte tydelige artefakter, eksempelvis i form af unaturlige ansigtsbevægelser, inkonsistent belysning eller synlige overgange mellem ansigt og baggrund. Moderne generative modeller er imidlertid langt mere sofistikerede, hvilket har øget behovet for avancerede maskinlæringsmodeller, der kan lære komplekse, ikke-lineære repræsentationer af billed- og videodata (Goodfellow et al., 2016, s. 530–533).

Konvolutionelle neurale netværk (CNN'er) er særligt velegnede til denne opgave, fordi de kan udnytte den rumlige struktur i billeddata og lære hierarkiske repræsentationer, hvor lavere lag identificerer simple mønstre som kanter og teksturer, mens højere lag fanger mere komplekse ansigts- og objektrelaterede strukturer (LeCun, Bengio & Hinton, 2015, s. 438–441).

5.4 Opbygning af et konvolutionelt neuralt netværk (CNN)

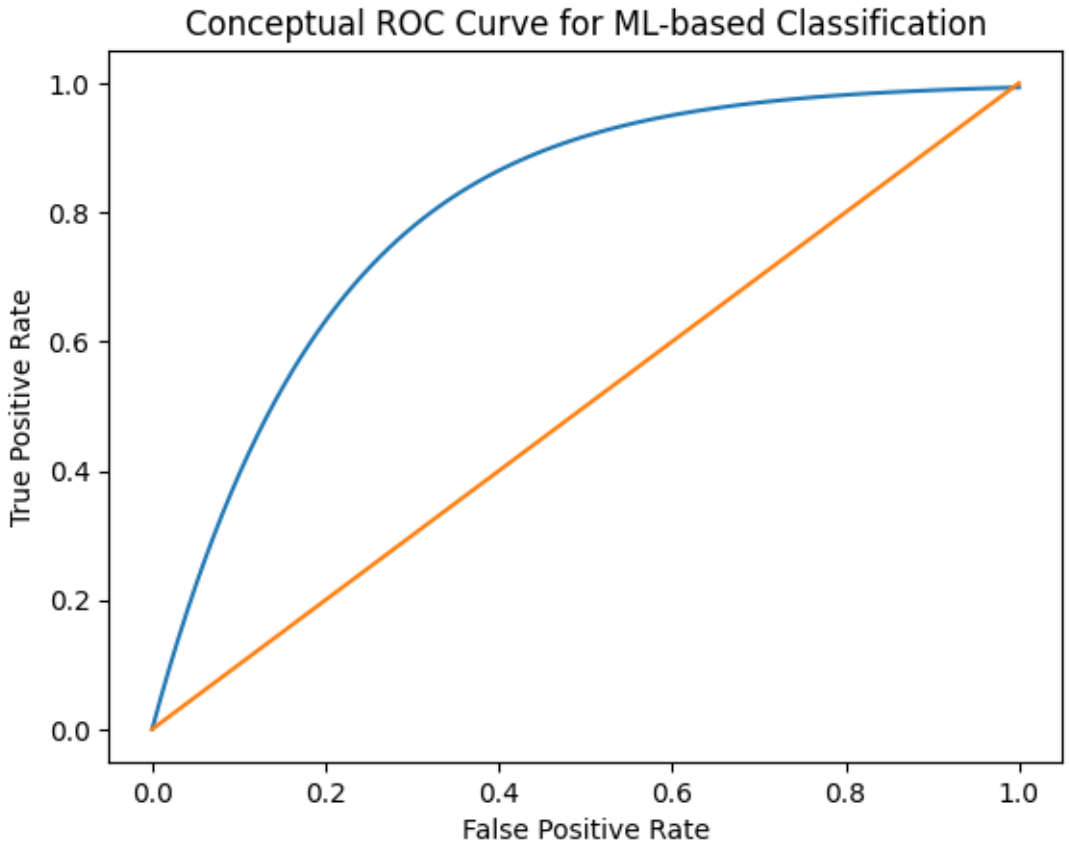
For at konkretisere anvendelsen af supervised learning i deepfake-detektion er der i dette speciale opbygget et konvolutionelt neuralt netværk baseret på billedframes udtrukket fra video. Modellen er implementeret i Python ved hjælp af PyTorch og trænet på et datasæt bestående af autentiske og manipulerede ansigtsbilleder (Paszke et al., 2019, s. 2–4).

CNN-arkitekturen består af en række gentagne konvolutionsblokke, hvor hver blok indeholder konvolutionslag, batch-normalisering, ikke-lineære aktiveringsfunktioner og pooling-lag. Denne struktur muliggør gradvis reduktion af billedets rumlige dimensioner samtidig med, at der opbygges mere abstrakte feature-repræsentationer (Goodfellow et al., 2016, s. 330–335).

Modellens output er en enkelt sandsynlighedsværdi, der angiver, hvorvidt et givent input klassificeres som deepfake. Træningen foregår ved at minimere en binær krydsentropi-lossfunktion, som straffer fejlagtige klassifikationer. Modellen evalueres ved hjælp af klassifikationsnøjagtighed og ROC-AUC, hvor sidstnævnte er særligt relevant i en valgsikkerhedskontekst, fordi den muliggør justering af beslutningstærsklen afhængigt af, om man ønsker at minimere falske positive eller falske negative (Fawcett, 2006, s. 862–864).

5.5 Modelperformance som normativt beslutningsrum

Denne tekniske demonstration illustrerer, hvordan deepfake-detektion i praksis beror på statistiske sandsynligheder snarere end sikre afgørelser. Modellen producerer ikke et definitivt svar på, om et medie er falsk, men en sandsynlighed, der kræver fortolkning og kontekstuel vurdering (Burrell, 2016, s. 3–6).



Figur 5.2: Konceptuel ROC-kurve for machine learning-baseret klassifikation i valgsikkerhed.

Figuren illustrerer forholdet mellem true positive rate og false positive rate for en binær klassifikationsmodel, f.eks. anvendt til deepfake-detektion eller DDoS-beskyttelse. Kurven viser, hvordan valg af beslutningstærskel indebærer normative afvejninger mellem sikkerhed og adgang, idet en lavere tærskel reducerer risikoen for falske negativer (under-beskyttelse), men øger risikoen for falske positiver (over-blocking af legitim adfærd).

Figur 5.2 anvendes i specialet som et analytisk redskab snarere end som et mål for modelperformance. ROC-kurven synliggør, at klassifikationsbeslutninger i maskinlæringsbaserede sikkerhedssystemer ikke er binære eller neutrale, men afhænger af valg af beslutningstærskel. En høj prioritering af true positive rate reducerer risikoen for, at angreb forbliver uopdagede, men øger

samtidig sandsynligheden for false positives, hvor legitim trafik eller indhold fejlagtigt klassificeres som ondsindet.

Valget af tærskel repræsenterer dermed ikke blot en teknisk justering, men en governance-beslutning, hvor balancen mellem sikkerhed, proportionalitet og adgang fastlægges gennem algoritmisk design. Dette understøtter Rudins pointe om, at høj prædiktiv nøjagtighed ikke i sig selv er tilstrækkelig i normative domæner, hvor fejlkostnader er asymmetriske og politisk betydningsfulde (Rudin, 2019). Samtidig illustrerer figuren Yeungs argument om algoritmisk regulering, hvor regulerende valg indlejres i tekniske systemer fremfor eksplicitte politiske processer (Yeung, 2018).

ROC-AUC fremstår derimod som en mere governance-relevant metrik, fordi den ikke fastlåser modellen til én beslutningstærskel, men synliggør et beslutningsrum, hvor valgmyndigheder og sikkerhedsaktører aktivt må afveje beskyttelse mod adgang. Dermed bliver valget af tærskel et politisk-administrativt valg snarere end et rent teknisk parameter.

False positives og false negatives repræsenterer i denne sammenhæng to forskellige demokratiske risici. Hvor falske negativer indebærer risiko for uopdagede angreb og dermed svækkelse af output-legitimiteten, indebærer falske positiver risiko for over-blocking, hvilket kan underminere proportionalitet, fairness og borgernes adgang til information og digitale tjenester. Fejlprofiler bliver dermed centrale for vurderingen af maskinlæringens demokratiske konsekvenser.

For at synliggøre, hvordan tekniske præstationsmål i maskinlæringsmodeller oversættes til demokratiske implikationer, præsenterer Tabel 5.2 en systematisk kobling mellem centrale ML-metrikker, deres tekniske fortolkning og deres demokratiske relevans i en valgkontekst.

Metrik	*Fortolkning*	*Demokratisk relevans*
Accuracy	Samlet korrekthed	Utilstrækkelig alene
ROC-AUC	Robusthed	Threshold-fleksibilitet
False positives	Fejlklassificeret legitim trafik	Risiko for over-blocking
False negatives	Oversete angreb	Risiko for systemisk sårbarhed

Tabel 5.2: Modelperformance og fejltyper i en demokratisk kontekst

Tabellen viser, at traditionelle performance-metrikker ikke er normativt neutrale. Særligt accuracy fremstår som et utilstrækkeligt mål i en valgkontekst, idet høj samlet korrekthed kan dække over alvorlige fejltyper med asymmetriske konsekvenser. En model kan f.eks. opnå meget høj accuracy samtidig med, at en relativt lille gruppe af legitime aktører systematisk fejlagtigt afskæres.

5.6 Sammenligning af modeller: Forklarbarhed versus teknisk effektivitet

For yderligere at konkretisere disse afvejninger sammenlignes i Tabel 5.3 to forskellige modeltyper anvendt i specialets tekniske analyse: en forklarlig baseline-model (logistisk regression) og en kompleks 1D-konvolutionel neuralt netværk.

Dimension	*Logistisk regression*	*1D-CNN*
Forklarbarhed	Høj	Lav-moderat
False positives	Højere	Lavere
False negatives	Lav	Meget lav
Kompleksitet	Lav	Høj
Governance-egnethed	Høj transparens	Høj effektivitet

Tabel 5.3: Sammenligning af logistisk regression og 1D-CNN

Tabellen tydeliggør, at valget af model indebærer et grundlæggende trade-off mellem forklarbarhed og teknisk performance. Den logistiske regression tilbyder høj grad af transparens og efterprøvelighed, hvilket styrker throughput-legitimiteten og gør modellen velegnet i kontekster med stærke krav om ansvarlighed og forklarbarhed. Til gengæld medfører den en højere false positive rate, hvilket øger risikoen for over-blocking.

1D-CNN-modellen udviser derimod markant bedre teknisk performance, særligt i forhold til at reducere både falske negativer og falske positiver. Dette styrker output-legitimiteten ved at sikre mere stabil og robust beskyttelse mod angreb. Samtidig reducerer modellens kompleksitet og begrænsede forklarbarhed mulighederne for offentlig indsigt og demokratisk kontrol.

Sammenligningen understøtter specialets centrale argument: Valget af maskinlæringsmodel er ikke et neutralt teknisk valg, men en governance-beslutning, hvor forskellige former for legitimitet prioriteres forskelligt. Hvor forklarlige modeller favoriserer transparens og ansvarlighed, favoriserer komplekse modeller effektivitet og sikkerhed. Denne afvejning kan ikke løses teknisk alene, men kræver politisk og administrativ stillingtagen.

5.7 Modellens begrænsninger og tekniske sårbarheder

Selvom CNN-baserede modeller kan opnå meget høj performance under kontrollerede betingelser, er de forbundet med en række væsentlige begrænsninger, som er særligt relevante i en valgkontekst. For det første er modellernes præstation stærkt afhængig af træningsdataenes kvalitet og repræsentativitet. Hvis datasættet ikke afspejler variationer i eksempelvis opløsning, komprimering, belysning, sprogbrug eller ansigtstyper, risikerer modellen at generalisere dårligt og dermed producere systematiske fejl (Raji et al., 2020, s. 11–14). I en demokratisk kontekst kan sådanne skævheder føre til uforholdsmæssig påvirkning af bestemte grupper eller kommunikationsformer.

For det andet er både deepfake-detektion og ML-baseret cybersikkerhed præget af en vedvarende kapløbssituation mellem angreb og forsvar. Når nye detektionsmetoder introduceres, tilpasses generative modeller og angrebsteknikker løbende for at omgå dem (Chesney & Citron, 2019, s. 151–154). Dette indebærer, at maskinlæringsbaserede sikkerhedssystemer ikke kan forstås som permanente løsninger, men snarere som midlertidige stabiliseringer i et dynamisk trusselslandskab.

Den tekniske opbygning af CNN-modellen tydeliggør samtidig, hvordan machine learning i valgsikkerhed indebærer en form for automatiseret vurdering af informationsautenticitet og legitim digital adfærd. Denne automatisering kan styrke output-legitimiteten ved at muliggøre hurtig og

effektiv håndtering af store datamængder, men den introducerer også en øget afhængighed af komplekse og delvist uigennemsigtige teknologiske infrastrukturer (Zerilli et al., 2019, s. 555–578).

Et centralt problem i denne sammenhæng er modellernes begrænsede forklarbarhed. Selvom teknikker som Grad-CAM kan give indblik i, hvilke datasegmenter eller billedområder modellen vægter i sine beslutninger, forbliver den samlede beslutningsproces vanskelig at forklare på en måde, der er tilgængelig for ikke-tekniske beslutningstagere (Selvaraju et al., 2017, s. 618–622). Dette udfordrer mulighederne for gennemsigtighed, ansvarlighed og efterprøvelse, som er centrale elementer i demokratisk styring.

Sammenligningen mellem den forklarlige baseline-model (logistisk regression) fra afsnit og den komplekse 1D-CNN-model illustrerer denne problematik yderligere.

Metrik	**Logistisk regression (baseline)**	**1D-CNN (kompleks model)**
AUC	0.9993	0.9998
False positive rate (FPR)	0.0647	0.0045
False negative rate (FNR)	0.0009	0.0000
Nøjagtighed	0.9983	0.9999

Tabel 5.4: Sammenligning af modelperformance og fejlprofiler

Som vist i Tabel 5.4 opnår CNN-modellen en markant lavere false positive rate og eliminerer falske negativer, hvilket styrker beskyttelsen mod uopdagede angreb og dermed output-legitimiteten. Den logistiske regression udviser derimod en højere false positive rate, hvilket indebærer en større risiko for over-blocking, men tilbyder samtidig et mere transparent beslutningsgrundlag. CNN-modellen reducerer denne risiko markant og eliminerer samtidig falske negativer, hvilket styrker beskyttelsen mod uopdagede angreb.

Samlet set viser analysen, at maskinlæringsbaserede sikkerhedssystemer opererer gennem probabilistisk styring. Beslutninger om, hvad der betragtes som legitim information eller legitim digital adfærd, træffes på baggrund af sandsynligheder og tærskelværdier fremfor entydige juridiske kriterier. Dette har direkte demokratiske implikationer: Mens høj teknisk effektivitet kan styrke valgsikkerheden, kan manglende gennemsigtighed og asymmetriske fejlprofiler udfordre både proportionalitet og demokratisk ansvarlighed.

Disse erkendelser danner et centralt analytisk udgangspunkt for de efterfølgende caseanalyser, hvor det undersøges, hvordan maskinlæring i praksis indgår i beskyttelsen af europæiske valg, og hvordan teknologiske valg omsættes til konkrete konsekvenser for politisk legitimitet, borgernes tillid og det demokratiske råderum.

Modellerne fungerer således ikke blot som tekniske værktøjer, men som styringsmekanismer, der institutionaliserer normative valg i teknologiske infrastrukturer. Denne erkendelse danner et centralt analytisk grundlag for de efterfølgende caseanalyser og for specialets samlede konklusion om maskinlæring som styringsteknologi i demokratiske systemer.

6. CASEANALYSE 1

Deepfakes, misinformation og demokratisk legitimitet ved parlamentsvalget i Slovakiet 2023

6.1 Kontekst: Valg, informationsmiljø og teknologisk sårbarhed

Parlamentsvalget i Slovakiet i september 2023 fandt sted i et mediemiljø præget af høj politisk polarisering, lav tillid til institutioner og udbredt brug af sociale medier som primær informationskilde. I denne kontekst blev generativ kunstig intelligens et centralt redskab i politisk kommunikation og manipulation. Som Bohacek fremhæver, var Slovakiet et af de første europæiske eksempler på, at AI-genererede deepfakes blev anvendt direkte i en valgkamp med det formål at påvirke vælgernes opfattelser (Bohacek, M.; and Farid, H. 2022, s. 1).

Det digitale informationsmiljø var karakteriseret ved høj hastighed, lav redaktionel filtrering og begrænset institutionel kapacitet til realtidsverifikation. Dette skabte gunstige betingelser for spredning af manipuleret indhold og forstærkede den epistemiske sårbarhed, som Vainaite betegner som en central udfordring for moderne demokratier (Vainaite, 2025, s. 2).

6.2 Deepfake-hændelsen og dens politiske betydning

Den mest omtalte hændelse under valgkampen var en lyddeepfake, der foregav at vise oppositionslederen Michal Šimečka i en samtale om valgmanipulation. Optagelsen blev offentliggjort kort før valget og spredt bredt på sociale medier. Ifølge Bohacek var timingen afgørende: deepfaken blev lanceret på et tidspunkt, hvor faktatjek og institutionel respons ikke kunne nå at korrigere narrativet inden stemmeafgivningen (Bohacek, M.; and Farid, H. 2022, s. 1–2).

Deepfaken havde ikke nødvendigvis til formål at overbevise vælgerne om en konkret påstand, men snarere at skabe tvivl, forvirring og mistillid. Dette understøtter teorien om liar's dividend, hvor eksistensen af manipuleret indhold gør det muligt at så tvivl om både falske og autentiske optagelser. Effekten er en generel erosion af tilliden til politisk kommunikation snarere end en entydig holdningsændring.

6.3 Machine learning og deepfake-detektion: Potentiale og begrænsninger

For at konkretisere de ellers abstrakte diskussioner om AI-baseret misinformation har specialet anvendt en konvolutionel neuralt netværksmodel til klassifikation af syntetisk og autentisk lyd. Modellens resultater viser en høj samlet klassifikationsperformance, herunder en høj AUC-værdi og en meget lav false negative rate.

Disse resultater indikerer, at machine learning i princippet kan fungere som et effektivt teknisk værktøj til identifikation af deepfake-indhold. Samtidig synliggør modellen også væsentlige begrænsninger. En ikke-triviel mængde false positives indebærer, at autentisk indhold risikerer at blive fejlagtigt klassificeret som manipuleret. I en valgkontekst er dette ikke blot et teknisk problem, men et demokratisk dilemma, da fejlagtig mistænkeliggørelse af legitim politisk kommunikation kan skade ytringsfrihed og politisk deltagelse.

Modellens output understreger dermed Rudins pointe om, at høj prædiktiv performance ikke nødvendigvis medfører høj demokratisk legitimitet, særligt når modeller anvendes i normative domæner præget af kontekst og fortolkning (Rudin, 2019, s. 206–210).

6.4 Institutionel fravær og governance-udfordringer

En central observation i Slovakiet-casen er det institutionelle fravær af machine learning i den officielle håndtering af deepfake-truslen. Deepfaken blev identificeret og problematiseret primært gennem journalister, uafhængige eksperter og efterfølgende platformstiltag. Der eksisterede ikke en formaliseret, offentlig forankret mekanisme til realtidsdetektion eller respons.

Vainaite beskriver denne type respons som reaktiv og fragmenteret, hvor staten i praksis overlader ansvaret for epistemisk sikkerhed til private platforme og medieaktører (Vainaite, 2025, s. 4–6). Dette rejser et centralt governance-problem: Når beslutninger om, hvad der er autentisk eller manipuleret, træffes uden klare institutionelle procedurer, svækkes både ansvarlighed og demokratisk kontrol.

6.5 Analyse: Legitimitet og epistemisk sikkerhed

Anvendt på den tredelte legitimitetsmodel rammer Slovakiet-casen især input-legitimiteten. Deepfake-indholdet forurenede vælgernes informationsgrundlag og underminerede forudsætningen for informerede valg. Når vælgere ikke kan skelne mellem autentisk og manipuleret kommunikation, svækkes den normative antagelse om rationel og informeret deltagelse.

Throughput-legitimiteten blev ligeledes udfordret af fraværet af klare procedurer og gennemsigtighed i håndteringen af misinformation. Borgerne havde begrænset indsigt i, hvem der traf beslutninger om indholdsmoderation, og på hvilket grundlag disse beslutninger blev truffet.

Output-legitimiteten fremstår mere ambivalent. Selve valgafviklingen blev gennemført korrekt, men valgets overordnede integritet blev udfordret af den udbredte tvivl og mistillid, som deepfaken bidrog til at skabe. Dermed illustrerer casen, at korrekt procedure ikke nødvendigvis er tilstrækkelig til at sikre demokratisk legitimitet, hvis det epistemiske grundlag er undermineret.

6.6 Delkonklusion

Slovakiet-casen viser, hvordan generativ AI og deepfakes kan udgøre en alvorlig demokratisk udfordring, selv uden teknisk indgriben i selve valgsystemerne. Fraværet af institutionelt forankrede ML-baserede detektionssystemer betød, at håndteringen af manipulation blev reaktiv, fragmenteret og i høj grad overladt til private aktører.

Specialets tekniske model demonstrerer, at machine learning i princippet kan identificere deepfake-indhold med høj præcision, men også at teknologien indebærer betydelige risici for fejlklassifikation og manglende forklarbarhed. I en valgkontekst gør dette fuldautomatiseret anvendelse problematisk.

Casen understøtter dermed specialets overordnede argument: Machine learning kan ikke betragtes som en neutral løsning på misinformation, men må forstås som en styringsteknologi, hvis demokratiske konsekvenser afhænger af institutionel indlejring, governance-rammer og normative afvejninger.

Hvor Case 1 illustrerer konsekvenserne af manglende institutionel anvendelse af machine learning i informationsdomænet, analyserer kapitel 7 en kontrasterende situation, hvor ML-baserede systemer er dybt integreret i beskyttelsen af den digitale valginfrastruktur. Overgangen fra deepfakes til DDoS synliggør dermed, hvordan trusselskarakteren former både teknologiens anvendelse og dens demokratiske implikationer.

7. CASEANALYSE 2

Cyberangreb og infrastrukturel beskyttelse ved valget til Europa-Parlamentet i 2024

7.1 Kontekst: Digital valginfrastruktur og trusselsbilledet

Valget til Europa-Parlamentet i juni 2024 fandt sted i en sikkerhedspolitisk kontekst præget af krigen i Ukraine og et generelt eskalerende cybertrusselsniveau mod europæiske institutioner. Som Saeva og Tasheva påpeger, udgjorde valget en *"critical vote"* i en periode, hvor både klassiske cyberangreb og hybride trusler udfordrede tilliden til og integriteten af demokratiske processer (Saeva & Tasheva, 2024, s. 226–227).

EU-valgets digitale infrastruktur er karakteriseret ved organisatorisk fragmentering på tværs af 27 medlemsstater, men samtidig ved fælles afhængighed af globale internetfunktioner såsom DNS, BGP og cloud-baserede platforme. Disse funktioner udgør det, Broeders og Sukumar betegner som *the public core of the internet*, som muliggør basal netværksfunktionalitet på tværs af nationale grænser (Broeders, D., & Sukumar, A., 2024, s. 411–414). Angreb mod denne kerne kan derfor få konsekvenser, der rækker langt ud over den enkelte stats valgsystem.

Parallelt hermed fandt valget sted i et informationsmiljø præget af udbredt anvendelse af generativ AI. CETaS dokumenterer, hvordan automatiserede bot-netværk og AI-genereret indhold i stigende grad blev anvendt til at forstærke politiske narrativer og skabe indtryk af bred folkelig opbakning (Stockwell et al., 2024, s. 8–16). Trusselsbilledet var således todelt: et infrastrukturelt lag,

hvor tekniske systemer kunne angribes, og et kognitivt lag, hvor vælgernes opfattelser kunne påvirkes.

7.2 Cyberangreb under EU-valget 2024

Under valgperioden i juni 2024 blev en række europæiske lande ramt af koordinerede DDoS-angreb rettet mod politiske partier, valgmyndigheder og offentlige informationsportaler. Ifølge Saeva og Tasheva fordobledes antallet af hacktivistangreb fra fjerde kvartal 2023 til første kvartal 2024, hvilket indikerer en systematisk kapacitetsopbygning frem mod valget (Saeva & Tasheva, 2024, s. 227–228).

Formålet med angrebene var ikke at kompromittere stemmeoptælling eller valgresultater direkte, men at gøre centrale kommunikationskanaler utilgængelige og dermed skabe usikkerhed om myndighedernes kontrol og beredskab. Samtidig identificerede CETaS en bølge af AI-baserede influence operations, herunder astroturfing og voter targeting via sociale medieplatforme. Over 40 TikTok-konti blev afsløret i at anvende AI text-to-speech til at sprede desinformation om valgsvindel og pro-Kremlin-narrativer om Ukrainekrigen (Stockwell et al., 2024, s. 16–17).

Kombinationen af infrastrukturelle angreb og kognitiv påvirkning udgjorde en klassisk hybrid trussel, hvor målet ikke var at ændre stemmer direkte, men at destabilisere det informationsmæssige og institutionelle grundlag for valget.

7.3 Teknologisk respons: Machine learning i DDoS-beskyttelse

På det infrastrukturelle niveau blev DDoS-angrebene i vidt omfang afbødet gennem ML-baserede netværksbeskyttelsessystemer, herunder cloud-baserede løsninger som Cloudflare. Disse systemer anvender unsupervised og semi-supervised learning til at identificere afvigelser i trafikmønstre, eksempelvis pludselige volumenspidser, atypiske forbindelser og geografiske uregelmæssigheder, og kan automatisk blokere mistænkelig trafik i realtid (Saeva & Tasheva, 2024, s. 227–229).

Specialets empiriske modellering understøtter denne beskrivelse. CNN-modellen opnåede meget lav false negative rate, hvilket indikerer høj effektivitet i forhold til at identificere og stoppe ondsindet trafik. Dette bidrager direkte til høj output-legitimitet gennem stabil drift af kritisk infrastruktur. Samtidig viser resultaterne, at false positives ikke kan elimineres fuldstændigt. Over-blocking indebærer, at legitim trafik i visse tilfælde fejlagtigt afskæres, hvilket kan have konsekvenser for adgang og proportionalitet.

Desuden viser analysen, at fejl ikke er jævnt fordelt på tværs af trafiktyper og subgrupper. Dette rejser spørgsmål om fairness og differentiel påvirkning i automatiserede sikkerhedssystemer og peger på, at selv teknisk effektive ML-løsninger indebærer normative valg.

På det kognitive niveau anvendte platforme som TikTok AI-baserede systemer til at identificere inauthentic behaviour, bot-netværk og AI-genereret indhold (TikTok, 2024, citeret i Saeva & Tasheva, 2024, s. 233). CETaS påpeger imidlertid, at mange influence operations undgik detektion ved at kombinere konventionelle bot-teknikker med generative værktøjer, hvilket gjorde dem mere adaptive og vanskelige at identificere (Stockwell et al., 2024, s. 15–17).

7.4 Institutionel forankring og governance-strukturer

I modsætning til Slovakiet-casen var EU-valget 2024 omgivet af et relativt formaliseret cybersikkerhedsberedskab. EU-institutioner, ENISA, nationale CERTs og private sikkerhedsleverandører indgik i et koordineret samarbejde med fokus på at sikre tilgængeligheden af offentlige kommunikationskanaler og beskytte den digitale infrastruktur (Saeva & Tasheva, 2024, s. 227).

Samtidig rejser Broeders og Sukumar et strukturelt governance-problem: Beskyttelsen af internettets public core er i høj grad afhængig af private aktører, mens internationale normer og retlige rammer for cyberoperationer fortsat er utilstrækkeligt udviklet (Broeders, D., & Sukumar, A., 2024, s. 412–423). Når valgsikkerhed i stigende grad hviler på private, proprietære ML-systemer, flyttes betydelig beslutningskompetence væk fra demokratisk kontrollerede institutioner.

Valget af klassifikationstærskler i ML-baserede DDoS-systemer fremstår i denne sammenhæng ikke som en neutral teknisk parameter, men som en governance-beslutning. Balancen mellem

beskyttelse og adgang fastlægges gennem teknisk design, hvilket indebærer, at centrale demokratiske afvejninger institutionaliseres i algoritmiske infrastrukturer med begrænset gennemsigtighed.

7.5 Analyse: Legitimitet, tillid og demokratisk råderum

Anvendt på den tredelte legitimitetsmodel påvirkede EU-valget 2024 input-, throughput- og output-legitimitet på forskellige måder. Input-legitimiteten blev indirekte udfordret af AI-baserede influence operations, der kunne skabe forvrængede opfattelser af politisk opbakning og sprede narrativer om valgsvindel (Stockwell et al., 2024, s. 15–17).

Throughput-legitimiteten blev på den ene side styrket af effektive, automatiserede DDoS-forsvar, men på den anden side udfordret af manglende gennemsigtighed i algoritmiske beslutningsprocesser. Borgernes mulighed for at forstå og efterprøve beslutninger om blokering og moderation er begrænset, hvilket kan svække oplevelsen af procesmæssig retfærdighed.

Output-legitimiteten fremstår relativt stærk, idet valget blev gennemført uden systemiske sammenbrud. Samtidig viser den empiriske analyse, at høj teknisk effektivitet ikke eliminerer normative spændinger, men snarere forskyder dem fra det politiske til det teknologiske niveau.

Det demokratiske råderum blev således både beskyttet og indskrænket: beskyttet mod teknisk sabotage, men udfordret af, at centrale beslutninger om legitim adgang træffes af automatiserede systemer uden direkte demokratisk ansvarlighed.

7.6 Delkonklusion

Case 2 viser, at machine learning kan anvendes effektivt til at beskytte valginfrastruktur mod infrastrukturelle cyberangreb og dermed bidrage til høj output-legitimitet. Samtidig illustrerer casen, at denne teknologiske effektivitet ikke er normativt neutral. ML-baserede sikkerhedssystemer indebærer valg om proportionalitet, fairness og ansvar, som i praksis institutionaliseres i teknologiske infrastrukturer.

EU-valget 2024 fremstår dermed som et teknologisk modent, men politisk komplekst eksempel på AI i valgsikkerhed. Hvor maskinlæring her faktisk beskytter den demokratiske proces mod teknisk forstyrrelse, introducerer den samtidig nye governance- og legitimitetsudfordringer, der ikke kan reduceres til spørgsmål om teknisk performance alene.

Samlet set viser Case 2, at machine learning i forbindelse med EU-valget i 2024 var institutionelt integreret som en central del af det cybersikkerhedsmæssige beredskab. De ML-baserede DDoS-beskyttelsessystemer bidrog til at sikre valginfrastrukturens funktionalitet og dermed en høj grad af output-legitimitet. Samtidig indebærer anvendelsen af automatiserede og i vidt omfang proprietære systemer, at afgørende beslutninger om legitim og illegitim netværkstrafik træffes uden fuld gennemsigtighed og uden direkte demokratisk ansvarlighed.

Denne situation adskiller sig markant fra Slovakiet-casen, hvor machine learning ikke var institutionelt operationaliseret i håndteringen af deepfake-baseret misinformation. Hvor EU-valget illustrerer konsekvenserne af en teknologisk moden og automatiseret valgsikkerhed, viser Slovakiet-casen konsekvenserne af fraværet af tilsvarende kapacitet i mødet med AI-baserede informationsangreb.

På baggrund af disse forskelle analyserer det følgende kapitel, hvordan anvendelsen – eller fraværet – af machine learning påvirker demokratisk legitimitet på tværs af forskellige trusselsdomæner. Gennem en komparativ analyse af informationsbaserede og infrastrukturelle angreb undersøges, hvordan maskinlæring både kan bidrage til beskyttelsen af demokratiske processer og samtidig rejse nye governance- og legitimitetsmæssige udfordringer.

8. KOMPARATIV ANALYSE

AI, cybersikkerhed og demokratisk legitimitet

De to cases repræsenterer fundamentalt forskellige måder, hvorpå kunstig intelligens og machine learning interagerer med demokratiske valgprocesser. Slovakiet-casen illustrerer en situation, hvor AI primært anvendes offensivt af angribere i form af deepfakes og informationsmanipulation, mens institutionelle aktører i begrænset omfang har operationaliseret teknologiske modforanstaltninger. EU-valget 2024 viser derimod et tilfælde, hvor machine learning er dybt integreret i den defensive infrastruktur, men hvor samme teknologiske logikker samtidig anvendes i influence operations.

Denne forskel synliggør en central spænding i analysen: Machine learning kan både fungere som beskyttelse af demokratiet og som en destabiliserende kraft. Samtidig viser casene, at selv velfungerende og teknisk effektive ML-systemer indebærer normative valg, der forskyder demokratiske afvejninger fra politiske institutioner til teknologiske infrastrukturer.

8.1 Trusselslogik: Informationsangreb versus infrastrukturelle angreb

En central forskel mellem de to cases ligger i truslernes karakter. I Slovakiet var den primære trussel epistemisk. Deepfake-indholdet rettet mod oppositionslederen Michal Šimečka havde til formål at skabe forvirring og mistillid på et afgørende tidspunkt i valgkampen. Som Bohacek fremhæver, var deepfaken designet til at skabe *"confusion and distrust at a decisive moment"* (Bohacek, M.; and Farid, H. 2022, s. 1). Dette understøtter Vainaites pointe om, at deepfakes udgør en trussel mod epistemic security, idet de underminerer borgernes mulighed for at skelne mellem autentisk og manipuleret information (Vainaitė, 2024, s. 2).

I EU-valget 2024 var trusselsbilledet mere komplekst. På den ene side blev den digitale infrastruktur udsat for DDoS-angreb og hacktivisme rettet mod valgrelaterede systemer og kommunikationskanaler (Saeva & Tasheva, 2024, s. 227–228). På den anden side blev vælgerne

udsat for AI-baserede influence operations, hvor bot-netværk og generativ AI forstærkede bestemte politiske narrativer, herunder påstande om valgsvindel og geopolitiske budskaber relateret til Ukrainekrigen (Stockwell et al., 2024, s. 16–17).

Hvor Slovakiet primært oplevede et kognitivt angreb, oplevede EU-valget således et hybridangreb, hvor både informationsmiljøet og den digitale infrastruktur var mål. Denne forskel har direkte implikationer for, hvordan machine learning kan og bliver anvendt.

8.2 Machine learning som fravær versus som infrastruktur

Den komparative analyse viser markante forskelle i den institutionelle forankring af machine learning. I Case 1 var anvendelsen af ML til deepfake-detektion ikke systematisk integreret i valgadministrationen. Identifikation og håndtering af manipulationen var i høj grad overladt til journalister, platforme og uafhængige eksperter. Som Vainaite påpeger, reagerede Slovakiet primært gennem efterfølgende remediation og adaptation, snarere end gennem teknologisk realtidsrespons (Vainaite, 2025, s. 4–6).

I Case 2 var machine learning derimod indlejret i den operative kerne af valgsikkerheden. Cloud-baserede ML-systemer klassificerede netværkstrafik i realtid og blokerede DDoS-angreb automatisk (Saeva & Tasheva, 2024, s. 227–229). Specialets empiriske resultater understøtter denne praksis ved at demonstrere meget lave false negative rates i DDoS-modellen, hvilket indikerer høj effektivitet i forhold til at beskytte infrastrukturens tilgængelighed.

Samtidig viser resultaterne, at false positives ikke kan elimineres fuldstændigt, og at fejl ikke er jævnt fordelt på tværs af subgrupper. Dette indikerer, at selv teknisk robuste ML-systemer indebærer fordelingsmæssige og normative konsekvenser, som sjældent synliggøres i den politiske debat om cybersikkerhed.

De to cases kan dermed forstås som yderpunkter på et strukturelt spektrum:

- Slovakiet - AI uden institutionel governance
- EU-valget - AI som infrastrukturel governance

8.3 Legitimitet på tværs af cases

Anvendt på den tredelte legitimitetsmodel viser analysen, at machine learning påvirker input-, throughput- og output-legitimitet forskelligt afhængigt af trusselsdomæne.

Input-legitimiteten blev mest direkte udfordret i Case 1. Deepfake-indholdet forurenede vælgernes informationsgrundlag og underminerede forudsætningen for informerede valg (Bohacek, M.; and Farid, H. 2022; Vainaite, 2025). I Case 2 var input-legitimiteten mindre direkte påvirket, idet stemmeafgivningen i sig selv ikke blev kompromitteret, men snarere omgivelserne for politisk kommunikation.

Throughput-legitimiteten var problematisk i begge cases, men af forskellige årsager. I Slovakiet skyldtes udfordringen fraværet af klare procedurer og institutionel kapacitet. I EU-valget skyldtes den derimod den teknologiske kompleksitet og manglende gennemsigtighed i algoritmiske beslutningsprocesser. I begge tilfælde vanskeliggør machine learning borgernes mulighed for at forstå og efterprøve centrale beslutninger. Output-legitimiteten fremstår stærkest i Case 2, hvor machine learning bidrog til stabil valgafvikling og høj teknisk robusthed. I Case 1 var output-legitimiteten mere ambivalent, idet valget teknisk blev gennemført korrekt, men oplevelsen af valgets integritet blev udfordret af epistemisk usikkerhed.

Analysen viser dermed, at høj output-legitimitet ikke nødvendigvis kompenserer for svækkelser i input- og throughput-legitimitet.

8.4 Liar's dividend og infrastrukturel afhængighed

Slovakiet-casen illustrerer det klassiske fænomen liar's dividend. Når deepfakes eksisterer, kan både ægte og falske optagelser afvises som manipulation, hvilket skaber en generel mistillid til medieindhold (Bohacek, M.; and Farid, H. 2022; Vaccari & Chadwick i Vainaite, 2020).

EU-valget 2024 illustrerer derimod en mere strukturel risiko: demokratisk infrastrukturel afhængighed. Beskyttelsen af valgprocessen bliver i stigende grad afhængig af private aktører, der kontrollerer centrale dele af internettets public core (Broeders, D., & Sukumar, A., 2024, s. 411–

423). Hvor Slovakiet lider under manglende teknologisk beskyttelse, lider EU-valget under afhængighed af uigennemsigtige teknologiske systemer.

I begge tilfælde udfordres borgernes tillid, men gennem forskellige mekanismer. I deepfake-casen sker det gennem synlig manipulation og epistemisk usikkerhed. I cyberangrebscasen sker det gennem erkendelsen af, at centrale demokratiske processer beskyttes af teknologier uden direkte demokratisk ansvarlighed.

8.5 Overordnet diskussion: AI, sikkerhed og demokratisk råderum

Den komparative analyse viser, at machine learning ikke er normativt neutral. Teknologien kan fungere som et effektivt værn mod tekniske angreb, men kan samtidig fungere som et demokratisk våben i informationsdomænet. Som Saeva og Tasheva fremhæver, kræver moderne valgsikkerhed både cyber-resilience og information security (Saeva & Tasheva, 2024, s. 226–227).

Samtidig understreger CETaS, at selv når AI ikke direkte ændrer valgresultater, skaber teknologien vedvarende epistemisk usikkerhed, som i sig selv kan være demokratisk destabiliserende (Stockwell et al., 2024, s. 8–9). Broeders og Sukumar påpeger desuden, at manglen på globale normer for beskyttelse af internettets public core skaber varige governance-huller i den internationale orden (Broeders, D., & Sukumar, A., 2024, s. 412–423).

Analysen peger dermed på behovet for differentierede governance-strategier. Machine learning bør anvendes forskelligt afhængigt af trusselskarakter og bør ledsages af klare krav til transparens, ansvarlighed og menneskelig kontrol.

På tværs af de to cases fremstår machine learning ikke blot som et teknisk redskab, men som en styringsteknologi. Algoritmiske systemer definerer i stigende grad, hvad der opfattes som legitim information og legitim adfærd, baseret på statistiske sandsynligheder snarere end juridiske kriterier.

Denne probabilistiske styring adskiller sig fra klassiske forvaltningsformer og udfordrer traditionelle forestillinger om ansvarlighed, gennemsigtighed og retssikkerhed. Samtidig indebærer den en

teknokratisering af valgsikkerheden, hvor centrale beslutninger flyttes ind i infrastrukturer, som er utilgængelige for offentlig indsigt.

Dermed opstår et grundlæggende demokratisk dilemma: Maskinlæring er nødvendig for at håndtere digitale trusler i stor skala, men kan samtidig underminere borgernes oplevelse af kontrol og deltagelse, hvis teknologien ikke forankres i klare institutionelle rammer.

8.7 Delkonklusion

Den komparative analyse viser, at machine learning spiller en afgørende, men ambivalent rolle i beskyttelsen af demokratiske valg. Hvor teknologien i infrastrukturelle sammenhænge bidrager til stabilitet og robusthed, er anvendelsen i informationsdomænet mere normativt kompleks og potentielt destabiliserende.

Specialet demonstrerer, at maskinlæring hverken er entydigt demokratifremmende eller demokratiundergravende. Teknologiens konsekvenser afhænger af dens institutionelle indlejring og de normative principper, der guider dens anvendelse. Dermed understøtter analysen specialets overordnede konklusion: Machine learning bør forstås som et politisk-administrativt fænomen, der kræver aktiv demokratisk styring snarere end som et neutralt teknisk værktøj.

8.7 Specialets forskningsbidrag

Dette speciale bidrager til den eksisterende forskning i krydsfeltet mellem kunstig intelligens, cybersikkerhed og demokratiske processer på flere niveauer – begrebsligt, empirisk, metodisk og policy-relevant.

For det første leverer specialet et begrebsligt bidrag ved at analysere maskinlæring som en styringsteknologi i demokratiske systemer. Hvor meget af den eksisterende litteratur enten behandler AI som et teknisk sikkerhedsværktøj eller som en normativ udfordring for demokratiet, viser dette speciale, at maskinlæring i valgkontekster fungerer som en form for politisk-administrativ infrastruktur. Gennem begreber som input-, throughput- og output-legitimitet samt algoritmisk governance tydeliggør analysen, hvordan AI ikke blot beskytter valg, men aktivt former,

hvad der opfattes som legitim information, legitim adfærd og legitim beslutningstagning. Dette perspektiv nuancerer en ofte polariseret debat, der enten fremstiller AI som løsning eller trussel, ved i stedet at fremhæve teknologiens dobbelte og kontekstafhængige rolle.

For det andet yder specialet et empirisk-analytisk bidrag gennem den komparative analyse af to forskellige typer digitale trusler mod europæiske valg: informationsbaserede trusler i form af deepfakes og infrastrukturelle trusler i form af cyberangreb. Ved at sammenholde Slovakiet 2023 og EU-valget 2024 demonstrerer specialet, at maskinlæringens funktion og demokratiske konsekvenser varierer markant afhængigt af trusselsdomæne. Hvor AI i deepfake-konteksten primært fungerer som en fraværende eller utilstrækkelig beskyttelse, fungerer den i infrastrukturel cybersikkerhed som et relativt modent og effektivt værn. Denne differentiering bidrager til en mere præcis forståelse af, hvor machine learning kan anvendes legitimt, og hvor teknologien risikerer at skabe nye demokratiske problemer.

For det tredje bidrager specialet metodisk ved at integrere teknisk modellering i en samfundsvidenskabelig analyse. Opbygningen af et konvolutionelt neuralt netværk til deepfake-detektion anvendes ikke som et rent ingeniørmæssigt projekt, men som et analytisk redskab til at synliggøre, hvordan maskinlæringsmodeller træffes beslutninger, hvilke typer fejl de producerer, og hvilke begrænsninger de har. Dette gør det muligt at koble abstrakte diskussioner om algoritmisk governance og legitimitet til konkrete tekniske mekanismer og dermed styrke den analytiske dybde.

Endelig udgør specialet et policy-relevant bidrag ved at identificere centrale governance-udfordringer i europæiske valgmyndigheders anvendelse af maskinlæring. Analysen peger på behovet for klare institutionelle rammer, transparenskrav, menneskelig kontrol og demokratisk ansvarlighed i brugen af AI til valgsikkerhed. Disse indsigter er direkte relevante for den fortsatte udvikling af europæisk regulering – herunder AI Act og Digital Services Act – og kan informere både fremtidig politisk beslutningstagning og praktisk implementering.

Samlet set bidrager specialet til forskningsfeltet ved at tilbyde en integreret analyse af maskinlæringens tekniske, institutionelle og demokratiske dimensioner i europæiske valgprocesser og ved at tydeliggøre de normative afvejninger, der følger med anvendelsen af AI i demokratiets beskyttelse.

9. KONKLUSION

Hvordan anvendes maskinlæring konkret af europæiske valgmyndigheder til at beskytte valgprocesser mod digitale trusler som misinformation, cyberangreb og deepfakes – og hvordan påvirker denne anvendelse politisk legitimitet, borgernes tillid og det demokratiske råderum i Europa?

Dette speciale har undersøgt, hvordan maskinlæring konkret anvendes i beskyttelsen af europæiske valgprocesser mod digitale trusler såsom misinformation, cyberangreb og deepfakes, samt hvilke implikationer denne anvendelse har for politisk legitimitet, borgernes tillid og det demokratiske råderum. Gennem en kombination af komparative casestudier, teknisk modellering og demokratiteoretisk analyse har specialet vist, at maskinlæring i stigende grad fungerer som en integreret del af demokratisk styring snarere end som et neutralt, teknisk sikkerhedsværktøj.

Analyserne af deepfake-hændelsen i Slovakiet i 2023 og cyberangrebene mod EU-valget i 2024 har demonstreret, at maskinlæring indgår direkte i udøvelsen af offentlig myndighed ved at klassificere information og adfærd som henholdsvis legitim eller illegitim. I deepfake-casen vedrører dette vurderingen af medieindholds autenticitet, mens det i cyberangrebscasen handler om, hvilken netværkstrafik der tillades eller blokeres. I begge tilfælde træffes afgørelserne på baggrund af statistiske sandsynligheder genereret af maskinlæringsmodeller snarere end eksplicitte juridiske eller politiske kriterier. Maskinlæring fungerer dermed som en form for probabilistisk styring, der adskiller sig grundlæggende fra klassiske forvaltningsformer baseret på regelbundne og begrundede beslutninger.

Denne styringsform har væsentlige demokratiske implikationer. Hvor traditionelle administrative afgørelser er forankret i transparens, forklarbarhed og mulighed for efterprøvelse, opererer maskinlæringsmodeller på baggrund af træningsdata og mønstergenkendelse, som ofte er vanskelige at gennemskue. Dette udfordrer centrale retsstatslige principper om ansvarlighed og borgernes mulighed for at forstå og anfægte beslutninger, der påvirker adgang til information og digitale infrastrukturer.

Specialets tekniske modellering har bidraget til at konkretisere disse abstrakte problemstillinger. Sammenligningen mellem en forklarlig baseline-model (logistisk regression) og en mere kompleks

1D-CNN-model viste, at begge modeller opnåede meget høj klassifikationsperformance målt ved AUC og nøjagtighed. Den væsentligste forskel fremkom imidlertid i fejlprofilerne. Den logistiske regression havde en markant højere false positive rate, hvilket indebærer en øget risiko for over-blocking af legitim trafik, mens CNN-modellen reducerede denne risiko betydeligt og samtidig eliminerede falske negativer. Disse resultater viser, at valget af model ikke blot er et teknisk spørgsmål om præcision, men et normativt valg med direkte konsekvenser for proportionalitet, adgang og fairness i automatiserede sikkerhedssystemer.

Modelresultaterne understøtter dermed analysens overordnede konklusion: at maskinlæring kan styrke output-legitimiteten ved at sikre stabil og robust valginfrastruktur, men samtidig kan udfordre throughput-legitimiteten, når afgørende beslutninger træffes af uigennemsigtige og proprietære systemer. Valget af klassifikationstærskler og tolerancen for henholdsvis falske positiver og falske negativer fremstår i denne sammenhæng ikke som neutrale tekniske parametre, men som governance-beslutninger, hvor balancen mellem sikkerhed og adgang fastlægges.

De to cases viser samtidig, at maskinlæringens demokratiske konsekvenser er kontekstafhængige. I Slovakiet blev input-legitimiteten undermineret af AI-genereret misinformation, mens fraværet af institutionelt forankrede ML-værktøjer førte til en fragmenteret og reaktiv håndtering. I EU-valget 2024 bidrog maskinlæring derimod til høj output-legitimitet gennem effektiv DDoS-beskyttelse, men skabte samtidig en strukturel afhængighed af private teknologileverandører og algoritmiske infrastrukturer, som borgerne har begrænset indsigt i. På tværs af begge cases viser specialet, at øget teknisk effektivitet ikke automatisk oversættes til øget demokratisk legitimitet.

Borgernes tillid fremstår i denne sammenhæng som både central og skrøbelig. I deepfake-casen udfordres tilliden gennem epistemisk usikkerhed og det såkaldte liar's dividend, hvor eksistensen af syntetisk indhold muliggør betvivlelse af både falske og autentiske informationer. I cyberangrebscasen opstår en mere latent tillidsproblematik, knyttet til bevidstheden om, at centrale demokratiske funktioner varetages af uigennemsigtige teknologiske systemer. Tillid bliver således i stigende grad medieret gennem teknologiske infrastrukturer fremfor direkte institutionel interaktion.

På denne baggrund konkluderer specialet, at maskinlæring i valgsikkerhed hverken er entydigt demokratifremmende eller demokratiundergravende. Teknologiens betydning afhænger af dens

institutionelle indlejring og de normative principper, der guider dens anvendelse. Maskinlæring bør derfor ikke fungere som erstatning for demokratisk styring, men som et supplement, der forudsætter klare rammer for ansvar, transparens og menneskelig kontrol. Beslutninger med politiske eller demokratiske konsekvenser bør ikke træffes fuldt automatiseret, men forankres i procedurer, der muliggør forklaring, efterprøvelse og offentlig debat.

Specialets samlede bidrag består i at integrere teknisk maskinlæringsanalyse med politologisk og forvaltningsmæssig teori. Ved at analysere konkrete modelresultater i sammenhæng med begreber om legitimitet og governance har specialet vist, hvordan maskinlæring aktivt former rammerne for demokratisk deltagelse og kontrol. Dermed bidrager specialet til en mere nuanceret forståelse af maskinlæring som styringsteknologi i demokratiske systemer og peger på behovet for en helhedsorienteret europæisk tilgang, hvor teknisk kapacitet ledsages af institutionel og demokratisk refleksion.

LITTERATURLISTE

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication, 33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Bohacek, M. 2023. Slovakia as the Precursor to Deepfake-Enabled Election Interference: Lessons Learned and Pathways Forward Matyas Bohacek. Stanford University, USA

Bohacek, M., & Farid, H. (2022). Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences, 119*(48), e2216035119. https://doi.org/10.1073/pnas.2216035119

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Broeders, D. (2024). Cybersecurity, elections and democratic legitimacy. I *Governing the digital domain* (pp. 1–15). Cambridge University Press.

Broeders, D., & Sukumar, A. (2024). Core concerns: The need for a governance framework for the public core of the internet. *Global Policy, 15*(3), 411–423. https://doi.org/10.1002/poi3.382

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Buzan, B., Wæver, O., & de Wilde, J. (1998). *Security: A new framework for analysis.* Lynne Rienner Publishers.

Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1819.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397.*

ENISA. (2024). *Threat landscape.* European Union Agency for Cybersecurity.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry, 12*(2), 219–245. https://doi.org/10.1177/1077800405284363

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences.* MIT Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review, 8*(4), 1–18. https://doi.org/10.14763/2019.4.1424

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Levi, M., & Stoker, L. (2000). Political trust and trustworthiness. *Annual Review of Political Science, 3*, 475–507. https://doi.org/10.1146/annurev.polisci.3.1.475

Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*, 8024–8035.

Raji, I. D., et al. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151. https://doi.org/10.1145/3375627.3375820

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saeva, T., & Tasheva, M. (2024). *The 2024 EU elections and cybersecurity: A retrospective and lessons learned.* ENISA. DOI:10.1177/17816858241288397

Scharpf, F. W. (1999). *Governing in Europe: Effective and democratic?* Oxford University Press.

Schmidt, V. A. (2013). Democracy and legitimacy in the European Union revisited. *Political Studies, 61*(1), 2–22. https://doi.org/10.1111/j.1467-9248.2012.00962.x

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual explanations from deep networks. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *2010 IEEE Symposium on Security and Privacy*, 305–316. https://doi.org/10.1109/SP.2010.25

Stockwell, S., Hughes, M., Swatton, P., Zhang, A., & Hall, J. (2024). *AI-enabled influence operations: Safeguarding future elections.* Centre for Emerging Technology and Security (CETaS). https://cetas.turing.ac.uk/publications

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation. *Social Media + Society, 6*(1), 1–13. https://doi.org/10.1177/2056305120903408

Vainaite, Vilija, Electoral Processes in EU Member States and Deepfake-based Disinformation: How do the Responses Differ? (August 30, 2024). Available at SSRN: https://ssrn.com/abstract=5039297

Yeung, K. (2018). Algorithmic regulation. *Regulation & Governance, 12*(4), 505–523. https://doi.org/10.1111/rego.12158

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. *Philosophy & Technology, 32*, 555–578. https://doi.org/10.1007/s13347-018-0332-6